UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

PROYECTO FIN DE CARRERA

Diseño y Análisis de Técnicas para la Mejora de la Caracterización de los Flujos de Red utilizando Muestreo Distribuido

Fernando Gutiérrez de Rubalcava Blanca

Junio de 2011

Diseño y Análisis de Técnicas para la Mejora de la Caracterización de los Flujos de Red utilizando Muestreo Distribuido

AUTOR: Fernando Gutiérrez de Rubalcava Blanca
TUTOR: José Luís García Dorado, PhD

High Performance Computing and Networking Research Group (HPCN)
Dpto. Tecnología Electrónica y Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Mayo de 2011

Resumen/Abstract:

Este proyecto está dedicado al estudio de sistemas de monitorización de red basados en Netflow muestreado. En él se estudia el impacto que el mismo impone en los datos obtenidos sobre una traza real y se evalúa una técnica que agrega las estadísticas obtenidas de manera distribuida en una red de datos, con buenos resultados.

This Project is dedicated to the study of Netflow-based monitorization systems. It studies the impact it causes on the data obtained from a real trace and evaluates a technique that aggregates the distributed statistics obtained in a data network, with good results.

Palabras Clave:

Monitorización de Redes, Flujos de Datos, Netflow, Identificación de Aplicaciones, Muestreo de Paquetes, Muestreo Distribuido, Chi Cuadrado.

Agradecimientos

A mis padres.

A todos aquellos seres queridos que, de un modo u otro, han compartido momentos buenos y malos conmigo durante este período en la Universidad, especialmente a mi hermana María, a Carla, a mis queridos amigos Jose Luís, Chevis y Santiago, a mis compañeros de clase y laboratorio, Cristina, Ana, Leticia, Sergio, Fernando, Alejandro, Eva, Borja, Mario… y muchos otros más que, como yo, han sabido muy bien en qué consiste experimentar la obstinación y determinación que implica querer aprender y conocer una carrera a la vez querida y complicada.

Gracias por estar ahí.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

1.1 *Motivación*

La monitorización del tráfico de Internet se ha convertido en una tarea fundamental para cualquier proveedor de servicio de Internet (ISP) así como en un tema de estudio y análisis para la comunidad investigadora.

El conocimiento y la monitorización del tráfico de datos en redes de comunicaciones proporcionan una valiosa información con numerosas aplicaciones y resulta, sin lugar a dudas, una tarea vital para cualquier proveedor de servicio, así como para otros miembros de la comunidad de Internet.

Por un lado, esta recolección permite a los proveedores de servicios de Internet (ISP) evaluar eficazmente las necesidades actuales y futuras de la red que gestionan y les aporta capacidad de decisión a la hora de valorar las soluciones a desplegar, en términos de arquitectura, tecnologías y equipos a utilizar que les permitan ofrecer calidad de servicio (QoS) y gestión de relación con clientes (CRM) apropiadas. En definitiva, esta información, si es suficientemente precisa, permite tener un control efectivo sobre las inversiones necesarias y la planificación general de la misma.

Por otro lado, la comunidad investigadora considera también esencial la captura de estas estadísticas de uso que permiten el estudio más avanzado de las dinámicas de Internet y el desarrollo de nuevas soluciones y modelos de red que beneficien a usuarios y proveedores.

En una Internet que cuenta cada día con mayor variedad de aplicaciones y servicios, con una demanda siempre creciente, con perfiles de utilización cada vez más imprevisibles, los ISPs se ven en la obligación de limitar estas inversiones y de actuar con precisión quirúrgica en las planificaciones mencionadas para mantenerse a la altura del nivel de competitividad existente.

Estos nuevos perfiles de uso, combinados con la necesidad de ofrecer soluciones económicamente atractivas para los clientes, dieron también lugar a nuevos tipos de ofertas

y modelos de tarificación, que hacen uso, de nuevo, de estos sistemas de monitorización, y que hacen aún más importante la efectiva utilización y elección de los mismos.

Aparte de estas consideraciones, la facilidad de acceso a las tecnologías de información genera también un alto nivel de vulnerabilidad, no solo para los usuarios en general, sino para los mismos proveedores, especialmente en lo que concierne a los ataques de denegación de servicios (DoS), cuyo objetivo es la "inundación" y posterior colapso de enlaces de red, que producen considerables pérdidas para estos. La detección, por lo tanto, de este tipo de ataques es otra importante aplicación de las medidas de tráfico de red.

Queda entonces clara la importancia de una monitorización efectiva. La elección, en consecuencia, de una métrica válida y robusta se tercia fundamental para esta tarea.

Algunos ejemplos de sistemas de monitorización pueden ser NetTraMet [1], Ntop [2], NG-MON [3], o aquellos basados en NetFlow (o flujos de red) [4], más tarde estandarizados por el Internet Engineering Task Force (IETF) a través de IPFIX [5], constituyendo estos últimos la herramienta elegida por la mayor parte de la industria a día de hoy . La medida fundamental establecida en estos es el flujo de red, o *NetFlow*. Este define una conexión de datos entre dos direcciones y puertos con un protocolo determinado. La implementación de esta definición paquete a paquete y la elaboración de listas de flujos ha sido el método empleado por la mayoría de equipos de red para obtener información del uso de la red en el tiempo.

El aumento, sin embargo de la velocidad de transmisión de las redes actuales ha comprometido el límite de escalabilidad para esta solución de monitorización mediante flujos de red, en concreto *Netflow*. Esto hace que el análisis paquete a paquete no sea posible, ya sea por los requerimientos de capacidad de procesamiento, velocidad de almacenamiento con memorias SRAM, sensiblemente más caras, o la propia sobrecarga introducida en la red para la exportación de grandes tablas de estadísticas.

Por esta razón, ambos grupos de trabajo del IETF, IPFIX [5] y PSAMP (Packet Sampling) han recomendado el uso de técnicas de muestreo de paquetes para resolver este problema [6]. De este modo, sólo un subconjunto de los paquetes que componen el tráfico a monitorizar son tenidos en cuenta para generar registros *Netflow*. Compañías como Cisco

utilizan muestreo estático del tipo “1 de cada k” paquetes en sus routers troncales de alta velocidad [7].

Este sistema, sin embargo introduce un cierto nivel de incertidumbre para determinadas aplicaciones y características de red que puede provocar que la bondad de sus resultados quede en algunos aspectos muy mermada. Ante este escenario, la comunidad científica se ha planteado el uso de técnicas de muestreo distribuido [8]. Esto es, el tráfico que circula por Internet es encaminado por múltiples routers, los cuales son capaces de generar estadísticas por cada flujo de red que los atraviesa. Sin embargo, la información que genera cada router es siempre analizada de forma aislada sin interacción alguna con el resto de la red. La motivación final de este proyecto fin de carrera es mostrar si efectivamente esta información “aislada” de cada enrutador puede ser agregada para construir unas estadísticas de red más precisas que las estadísticas generadas de forma independiente por cada router. En este sentido se pretende evaluar la mejora obtenida.

1.2 *Objetivos*

El presente Proyecto Fin de Carrera se centra en el estudio de la técnica de *NetFlow* muestreado, cuya adopción se considera ampliamente aceptada por operadoras de red y grupos de investigación, pero que cuenta con problemas inherentes en el nivel de precisión de sus medidas.

La finalidad de este proyecto, por lo tanto, pasa por arrojar mayor luz sobre el tema mediante una valoración empírica de esta técnica y la proposición, si procediera, de mejoras en la misma.

Los objetivos concretos de este proyecto son caracterizar a nivel de flujo el tráfico real de un operador, estudiar empíricamente el impacto que tiene el muestreo en la generación de los registros *Netflow* y analizar cómo el muestro distribuido puede mitigar este impacto.

Para llevar a cabo estos objetivos se pretende llevar a cabo un trabajo en distintas fases que implicarán entonces los siguientes puntos:

- Estudio del estado del arte en técnicas de adquisición de flujos de red.

Se atenderá a los resultados obtenidos por otros grupos para conocer la tendencia actual y se usarán como referencia para tomar decisiones a la hora de elegir parámetros, métricas, técnicas o cualquier otro aspecto que pueda concernir al Proyecto.

- Identificación de trazas reales de tráfico de Internet que permita evaluar empíricamente las mejoras obtenidas por el muestreo distribuido en un entorno real.

- Implementación de un programa que permita generar flujos de red tal y como lo hacen los routers actuales así como que permita nuevas técnicas de submuestreo como el muestreo distribuido.

- Estudio y análisis de las distintas métricas utilizadas para determinar la distorsión que añade el muestreo a las estadísticas de los flujos de red. Cabe destacar el reto que supone analizar trazas reales de red con alto nivel de agregación en términos computacionales. El empleo de grandes cantidades de datos acarrea problemas en la implementación del generador de flujos, en la lentitud de las ejecuciones y la dificultad para hacer una extracción de datos efectiva de sus salidas, que resultan a su vez un problema, no sólo de almacenamiento y tiempo, sino de diseño, al contar con limitaciones claras de memoria y potencia de proceso. De este modo se establece como objetivo de este trabajo la implementación de todo el código necesario de manera óptima prestando por tanto una especial atención al coste de ejecución.

- Comparativa del rendimiento del muestreo distribuido y otras técnicas.

- Propuesta, si procede, de una infraestructura para la implementación de un sistema de monitorización de flujos de red distribuido.

1.3 *Organización de la memoria*

La memoria se organiza en 5 capítulos.

El primer capítulo, Introducción, engloba este apartado, identifica la motivación de la realización de este Proyecto Fin de Carrera y expone sus objetivos y el planteamiento ideado para su conclusión.

En el segundo capítulo, Estado del Arte, se muestran primero los conceptos básicos necesarios para el estudio. A continuación y de manera ordenada,se explican las distintas ramas de investigación actuales que se consideran de relevancia para el estudio, y en las cuales se apoyarán ciertas decisiones y propuestas, tanto para el desarrollo del Proyecto, como para exponer posibles soluciones futuras basadas en los resultados obtenidos que puedan mejorar la adopción de sistemas Netflow.

El tercer capítulo, Caracterización del Tráfico, hace una descripción de la traza de datos utilizada como referencia para este estudio, y realiza una caracterización exhaustiva de la misma a través de distintas métricas, que son expuestas por primera vez en el estudio, y que servirán como referencia para los resultados obtenidos mediante el uso de distintas técnicas de muestreo, tanto individual como distribuido, en los capítulos posteriores. En este apartado se muestran también las herramientas desarrolladas y los medios utilizados para la conclusión del Proyecto.

El cuarto capítulo, Impacto del Muestreo, se divide a su vez en dos grandes apartados:
El primero, Muestreo convencional, realiza un estudio exhaustivo del impacto que tiene el muestreo de paquetes según distintos parámetros en los resultados obtenidos para cada métrica a considerar. Para ello se hace uso de tests de similitud usados como referencia en la literatura, cuyas definiciones y limitación son expuestas a la vista de los resultados y de su análisis.

El segundo, Muestro distribuido, propone el esquema de adopción de un sistema basado en *Netflows* muestreado y distribuido. A este efecto, y una vez realizadas las pruebas pertinentes mediante este sistema, se muestran de nuevo los resultados obtenidos según distintas métricas.

El último capítulo, Resultados, hace un análisis de todos los resultados obtenidos y muestra las conclusiones extraídas del estudio. También se proponen nuevas líneas de investigación que podrían llevarse a cabo a la luz de las mismas.

2 Estado del arte

2.1 *Netflows y muestreo*

2.1.1 Netflow

Un *Netflow* es una secuencia de paquetes que comparten las mismas direcciones IP, puertos origen y destino y el mismo protocolo [5]. Esta quíntupla es la que establece un flujo de datos en red según esta definición. La Figura 2-1 muestra en ejemplo concreto de una quíntupla Netflow.

Ilustración 2-1: Ejemplo de quíntupla de flujo de un paquete de datos

Normalmente un registro *Netflow* incluye, además de los ya nombrados IP origen, IP destino, Puerto Origen, Puerto Destino y Protocolo, el volumen agregado de trafico generado en bytes, número de paquetes total, índices de las interfaces de entrada y salida y marcas de tiempo de inicio y fin del flujo.

En la Figura 2-2 se muestra el funcionamiento general de un generador de flujos Netflow. Primero se ha de analizar el paquete extrayendo la quíntupla. Con este información se busca en la tabla de listado de flujos ya activos si el paquete que se está analizando pertenece a un flujo ya abierto. En caso contrario se crea un nuevo flujos en la tabla.

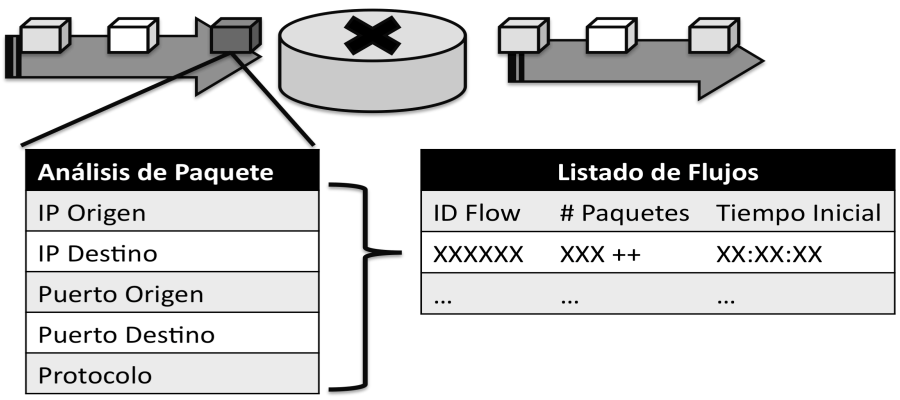


Ilustración 2-2: Proceso de identificación del flujo de un paquete

En un principio el sistema *Netflow* se implementó en los routers Cisco para reducir la carga de conmutación sobre los mismos. Este sistema de identificación de flujos permitía direccionar un paquete rápidamente sin tener que buscar en la tabla de direccionamiento, lo que implicaba una considerable reducción del coste computacional.

Más adelante se identificaron que las tablas de flujos activos eran buenas candidatas para recolectar datos de monitorización del uso de la red, por lo que se diseñó también un protocolo que permitía a los routers exportar esta información. Este protocolo se denominó también *Netflow*. El empleo de la palabra *Neftlow*, por lo tanto, no se limita a una definición de un flujo de red, sino también al informe consecuente del análisis de estos y al protocolo usado para su exportación a otros equipos.

Desde entonces, se han utilizado registros *Netflow* para distintas aplicaciones, como la facturación por uso, la obtención de perfiles de utilización, el dimensionamiento y planificación de red, o la detección de intrusiones y ataques de denegación de servicio [9].

La arquitectura de un sistema de monitorización basado en flujos cuenta con tres elementos principales de procesamiento: generador de flujos (figura 2-2), almacenador (o colector) de flujos y analizador de tráfico, como puede verse en la figura 2-3.

- El generador de flujos se sitúa en el punto o puntos de medida deseados en la red. Estos elementos se encargan de generar los informes correspondiente al tráfico que pasa por ellos.

- El almacenador es el punto de la red al que los distintos generadores exportan sus informes. Esta exportación es la que hace uso del ya mencionado protocolo *Netflow* proporcionado por Cisco. Este almacenamiento se puede llevar a cabo en distintos puntos según la necesidad, y es el paso previo al análisis del tráfico medido.

El analizador es la herramienta finalmente utilizada por el operador para procesar búsquedas de métricas o características concretas en los flujos de la red. La herramienta utilizada puede variar entre sistemas.

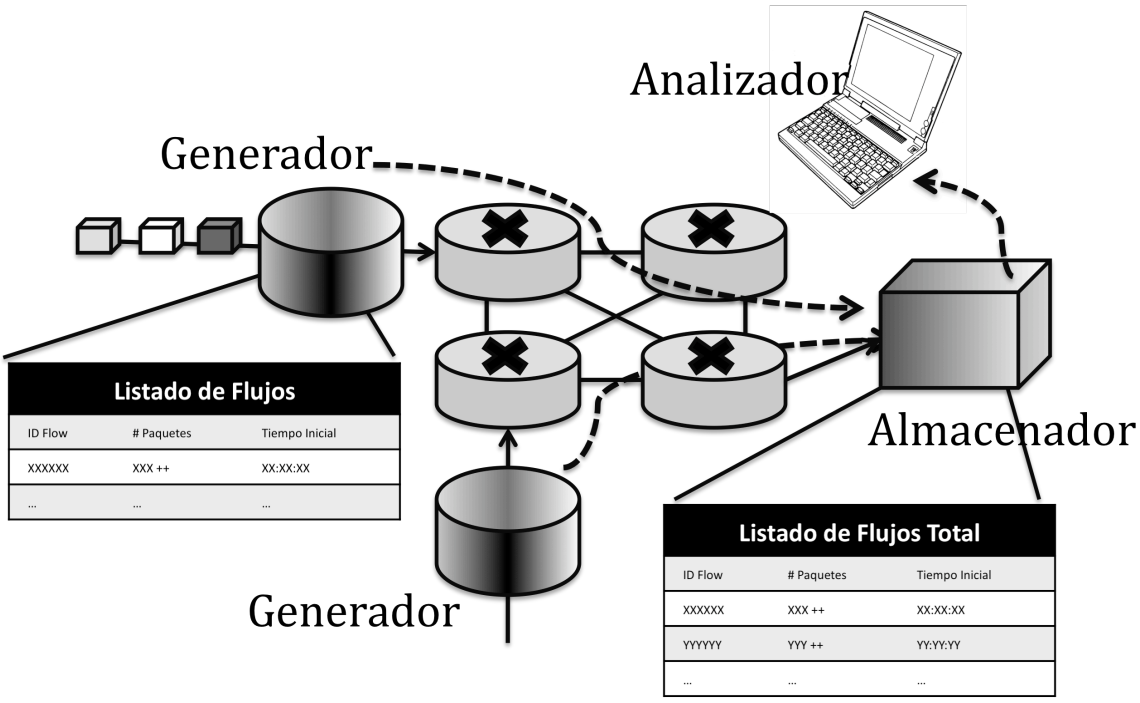


Ilustración 2-3: Esquema de funcionamiento de sistema de monitorización

Existen, sin embargo, limitaciones inherentes a la adopción de los *Netflows* como método de monitorización, tanto en su captura como en la creación de sus estadísticas.

Por un lado, un flujo según la definición inicial puede llegar a durar un tiempo ilimitado, en el rango de días o meses. Es claro que los equipos de red cuentan con una capacidad limitada de memoria, por lo que no pueden mantener tablas con grandes números de flujos abiertos a la espera de la llegada de un paquete nuevo. Para evitar esta situación se considera en general que un flujo ha terminado si se da alguna de las siguientes circunstancias:

- En el caso de una conexión TCP, la aparición de una bandera de finalización (FIN) o reset (RST).

- Si no se detecta tráfico en un flujo abierto durante un tiempo determinado. Existen distintos intervalos considerados en este caso.

- Cuando un flujo lleva abierto más de un período determinado, independientemente del trafico generado.

- Cuando la tabla de flujos del router está llena y es necesario espacio libre para añadir nuevas entradas.

Por otro lado, aparte de la limitación temporal a largo plazo, es hoy en día común que la intensidad del tráfico no permita el análisis de todos y cada uno de los paquetes y su inclusión en la tabla de flujos a tiempo antes de la llegada del siguiente. El factor principal de esta limitación se debe principalmente, según [10], al tiempo disponible entre llegadas. Es decir, la captura y lectura de cada paquete es más crítica que la generación de un flujo y su posterior escritura. Esto es dependiente, no sólo de la capacidad de proceso del equipo sino, y de manera crítica, de la velocidad de acceso y escritura en memoria. Estas últimas deben ser lo suficientemente rápidas como para mantener la velocidad de la línea. Esto determinaría hoy el uso de SRAM, lo que encarece en muchos casos sobremanera el diseño o incluso lo imposibilita debido a que el tamaño típico de la memoria SRAM es de unos pocos MB [11].

La solución adoptada por la industria ha sido el empleo de técnicas de muestreo de paquetes, de manera que sólo se utiliza un porcentaje del tráfico para hacer la estimación total.

2.1.2 Muestreo aleatorio vs. determinista

A la hora de realizar un muestreo del total de tráfico que pasa por un nodo de red existen dos alternativas básicas:

- **Muestreo aleatorio**, según el cual un paquete de llegada se analiza o se descarta según una probabilidad dada, p, por ejemplo mediante la generación de un número aleatorio en un rango y la aplicación de un umbral.
- **Muestreo determinista**, por el que se toma un paquete a cada k llegadas o eventos.

Se ha planteado la existencia de diferencias entre estos dos métodos fundamentales. En [12] se realiza una comprobación la precisión de ambos mediante comparación de sus resultados tras la ejecución frente a la misma traza de datos. Se hace un análisis teórico previo que compara la variabilidad de las estimaciones de tamaño de flujo que se harían

con uno y otro método. Se llega a la conclusión que la precisión es aproximadamente equivalente siempre que la población de paquetes que llegan al router esté ordenada aleatoriamente, es decir, no se repiten los paquetes de cada flujo con un periodicidad dada. El estudio pasa después por una parte empírica en la que se comparan los resultados de 200 ejecuciones de ambos métodos con una tasa de 1 de cada 250 para el caso determinista y una probabilidad de 1/250 para el aleatorio. El resultado obtenido concluye que las diferencias entre ambos son indistinguibles en la práctica.

Este estudio justifica que, a la hora de realizar nuestras propias simulaciones, nos centremosen los resultados obtenidos con muestreo determinista sin pérdida de generalidad al poder asumirlos equivalentes. Se elegirá muestreo determinista por ser el algoritmo de implementación más común en los equipos comerciales actuales y por permitir usar un cierto "offset" en la simulación sobre la misma traza de datos. Este offset representa el momento relativo decidido para comenzar a hacer captura. De este modo, y utilizando como ejemplo un muestreo de 1 paquete de cada 10 llegadas, se muestrean los paquetes 1, 11, 21, etc.… para un offset 0, y los 4, 24, 34, etc.… para un offset 3. Este concepto se explicará con más detalle en el tercer capítulo, y será utilizado para la propuesta de un esquema de muestreo distribuido.

2.1.3 Muestreo por tiempo vs. muestreo por evento

Las propuestas básicas de muestreo aleatorio y determinista pueden ser aplicadas a distintas líneas causales:

- Una **determinada por eventos,** en la que cada llegada de paquete induce una evaluación de captura o descarte.

- Una **línea temporal**, por la que cada evaluación se hace cada cierto intervalo de tiempo establecido.

Los autores de [13] tratan de comparar los métodos determinista y aleatorio antes nombrados en distintas condiciones, por evento y por tiempo. Sus resultados les llevan a desechar la opción de muestreo temporal por encontrar diferencias significativamente

mayores en las métricas realizadas con aquellas de la traza original, en comparación con aquellas obtenidas por muestreo por eventos, donde la diferencia era menor.

Otros grupos han probado sistemas híbridos. En [14] se prueba un sistema que realiza un muestreo temporal en casos de baja utilización del enlace, y muestreo por eventos para mayores tasas. Su investigación compara tres métodos: temporal, por eventos, e hibrido. Sus resultados evidencian de nuevo el peor rendimiento del temporal, frente a los otros dos, que obtienen resultados parecidos. La superioridad del método por llegada de paquetes parece clara a la luz de estos resultados y de su sencillez relativa frente a otros.

Estos resultados determinarán la adopción para nuestro experimento de muestreo determinista por llegada de paquetes, al probarse su valía frente a otros.

2.1.4 Otras técnicas de muestreo

Como alternativa al muestro de paquetes, existen estudios que han valorado el uso de otras técnicas, como el muestreo de flujos propuesto en [15].

El muestreo de flujos se basa en el uso de una función hash que otorgue un código único a cada una de las quíntuplasdistintivas de cada netflow: *IP origen, IP destino, Puerto Origen, Puerto Destino, Protocolo*. De este modo un router solo muestrea aquellos flujos contenidos en un rango determinado. La Figura 2-4 muestra esquemáticamente este sistema.

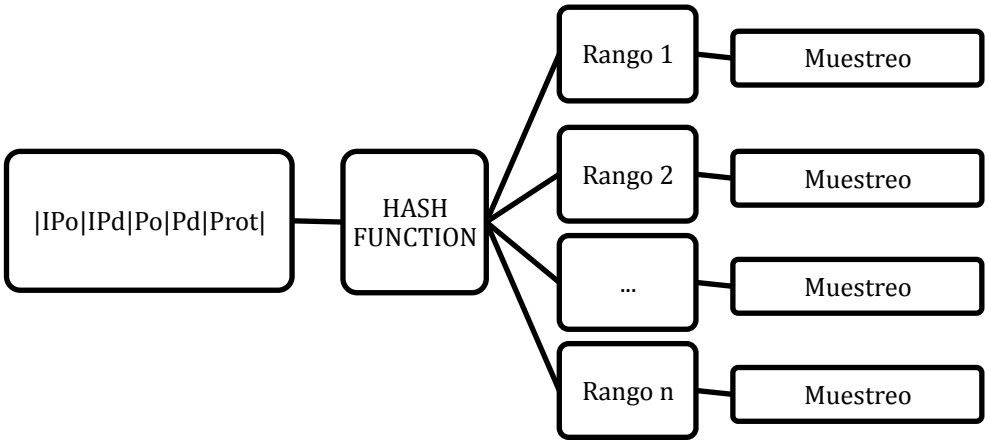


Ilustración 2-4: Distribución de rangos muestreo de flujos a través de función hash

En [16] se hace uso de esta técnica y propone la creación de un sistema que, mediante la computación de funciones *hash* sobre cada paquete que llega a cada router, haga una distribución del peso del muestreo entre equipos. Estas funciones *hash* son capaces de asignar una firma única a cada paquete que le relaciona de forma inequívoca con un flujo. De este modo, al monitorizar la red, sólo un conjunto de firmas se consideran de interés y únicamente un subconjunto de flujos se analizan. Este sistema estaría controlado de manera centralizada desde un punto de recolección por un sistema que determinaría qué flujos debe evaluar cada uno de los nodos, de modo que un flujo sea muestreado solo una vez en determinados puntos durante su transmisión dentro de la red de comunicaciones. La única comunicación explícita entre el centro de recolección y el resto sería el otorgamiento de la lista de flujos a muestrear.

Así, varios equipos cuentan con un rango concreto de flujos que deben muestrear. Este rango se denomina "*manifiesto*" y es establecido desde la central de control. El principal inconveniente de esta aproximación es que requiere un sistema dedicado para asignar rangos de la función *hash* así como flexibilidad ante cambios de rutas. El resultado mostrado por los autores muestra una cobertura efectiva de la mayoría del tráfico que pasa por la red si bien el escenario usado presenta ciertas restricciones.

El cálculo del código *hash* para su posterior evaluación obliga a que se analicen, en efecto, las cabeceras de todos y cada unos de los paquetes que llegan a cada router. Esto determina de nuevo el uso de memorias suficientemente rápidas como para soportar la velocidad de la línea, que es de hecho lo que la monitorización por muestreo de paquetes ha pretendido evitar desde su concepción. Si bien según la valoración hecha por el estudio, la cantidad de SRAM necesaria es relativamente poca, esta propuesta en cierto modo vuelve a topar con la limitación que fundamenta el uso del muestreo.

Los autores de [17] tratan de superar la necesidad de uso de memorias rápidas y propone un sistema que mezcla dos tipos de muestreo: muestreo de paquetes y muestreo de flujos. Este *muestreo doble* se realizaría según dos pasos:

- Un muestreo de aleatorio de los paquetes de llegada, es decir, con una probabilidad 1/n.

- Un muestreo del flujo identificado. Se computa el código hash y se acepta el paquete si pertenece a un rango establecido.

De este modo se logra disminuir la carga sobre los equipos a la vez que es posible establecer objetivos más claros en las estadísticas a obtener.

2.2 *Características y limitaciones.*

2.2.1 Nuevas características de tráfico.

Una de las características interesantes de la monitorización a través de flujos es la compresión de información que implica el paso de múltiples cabeceras de paquetes a una entrada de *Netflow*. Esto es especialmente obvio para flujos grandes, pero existe una nueva tendencia en el aumento de conexiones cortas que se mantienen por poco tiempo y que se reabren con puertos distintos, en contraposición con el tráfico "típico", que mantiene una conexión abierta durante la sesión y utiliza puertos bien conocidos para el establecimiento. Esto es aparentemente una característica común de servicios tipo P2P y en otros como gusanos y ataques DoS. Este tipo de tráfico genera muchos flujos distintos que cuentan con un número muy pequeño de paquetes, típicamente menor a 5. Según la administración de Internet2 [18],un 3% del tráfico total de su red pertenece a aplicaciones P2P, además de un 52% de tráfico restante no identificado que podría en buena medida deberse a lo mismo. En [10] los autores se hacen eco de esta característica y hacen un estudio del impacto de la misma. El documento confirma que este aumento del número de entradas *netflow* afecta gravemente a la cantidad de memoria necesaria en los equipos de red, y en menor medida a la carga en CPU. Existen investigaciones que intentaron superar este problema mediante exportación adaptativa de estadísticas de flujos [19], o la definición de un nuevo tipo de flujo [20].

De nuevo en el estudio de [10] se hace una valoración de este tipo de flujos cortos, denominados *flash flows*, que cumplen con tres características:

- Duración total menor que 1 segundo.
- Número de paquetes menor o igual a 3.
- Número total de Bytes menor que 500.

En las medidas utilizadas en este trabajo, más del 50% del número total de Netflows pertenece a este grupo.

Se considera por lo tanto, interesante el análisis de esta métrica en nuestro estudio al probarse una característica distintiva del tráfico actual, cuyo correcto estudio puede llevar a mejoras en los sistemas de monitorización.

2.2.2 Detección de anomalías.

En estudios como el realizado en [10] se pone de manifiesto cómo una fuente importante de flash flows encontrados en la Red está efectivamente asociada a servicios legítimos, como P2P, pero también, y en gran medida, a aplicaciones maliciosas como gusanos y ataques de denegación de servicio. Estos producen generalmente conexiones UDP en puertos no conocidos que corresponden a flujos muy cortos, como los *flash flows* antes definidos.

Los autores de [21] hacen un análisis del impacto que tiene el muestreo de paquetes en la detección de este tipo de anomalías. Se utiliza una traza de datos de una red que infectada con un gusano y se muestrea a diferentes tasas. Después se comparan las métricas obtenidas en cada caso. El hecho de que los flujos pequeños son capturados con mucha menor probabilidad da lugar a que, si bien las métricas de número de bytes y paquetes no se ven comparativamente muy afectadas, la del número de flujos puede llegar a un alto grado de enmascaramiento. El documento expone los resultados del uso de sumatorios basados en entropías del numero de paquetes y flujos, que es capaz en cierta medida de sobreponerse a ese enmascaramiento.

3 Caracterización de los Flujos de Red

3.1 Introducción

Esta sección muestra el proceso llevado a cabo para la caracterización de los datos sometidos a estudio mediante *netflows*.

Comienza con una descripción preliminar de la traza de datos seleccionada y continúa explicando los medios utilizados para su análisis. Inmediatamente después se describen las métricas tenidas en cuenta y el proceso llevado a cabo para su obtención. Por último se muestran los resultados obtenidos. Esto permite contar con una información importante a la hora de comprobar el impacto que el posterior muestreo de paquetes pueda tener en las estadísticas obtenidas.

3.2 Tráfico Analizado

La traza de datos elegida para este estudio corresponde a una recolección realizada en la red de un operador de red español a través de accesos 3G de usuarios móviles.

La traza comienza el 8 de Marzo de 2009 a las 23:00 GMT y acaba al día siguiente a las 20:14:03. El tiempo total de la traza, por tanto, es de aproximadamente 21 horas. El grueso total de datos extraídos en ese tiempo es de 27.6 Gigabytes.

El contenido de la misma está en formato *pcap*. Éste es un formato estándar, que utilizan de manera común programas de análisis de protocolos, como*Wireshark*[22]. A la hora de elaborar un código propio para su estudio, se hará uso de la librería *libpcap*[23]para la lectura y extracción de los campos de cada paquete.

3.3 Herramientas Implementadas y Empleadas

El proceso de obtención de datos comienza por la creación de un programa que sea capaz de simular la extracción de informes *Netflow* de la traza de datos original. Una vez obtenidos estos informes se extraerán ciertas estadísticas a través de otro software que permita realizar cálculos y gráficas de manera más accesible.

3.3.1 Generador de Informes Netflow

La extracción de informes *netflow* se realiza vía un generador de informes que sea capaz de procesar un archivo de datos *pcap* con información de todos los paquetes de la traza obtenida para análisis que pertenezcan a los protocolos TCP o UDP.

El generador finalmente implementado hace uso de la librería *libpcap* para esta tarea. Esta proporciona una serie de funciones para la captura y análisis de paquetes. En el caso de este estudio, los datos a analizar forman parte de una traza capturada anteriormente, por lo que se hace uso de una funcionalidad de apertura *offline*. Para la familiarización con esta librería se hizo uso de [24].

La entrada al programa vendrá dada obligatoriamentepor el archivo *pcap* a analizar, el tiempo máximo de flujo, el tiempo máximo entre llegadas, el tipo de muestreo y su tasa. Mediante estos parámetros éste es capaz de generar las listas de *netflows* existentes en la traza, sobre las que se recogerán una serie de métricas que se compararán en la fase siguiente del estudio.

3.3.1.1 Entorno de Trabajo

La simulación se realiza sobre un servidor que cuenta con un procesador Intel Xeon 64bits a 2.66GHzcon 4 Gigas de RAM disponibles sobre el que corre un Sistema Operativo *Ubuntu* versión 10.04.1 “Lucid” con kernel 2.6.32. La comunicación con el mismo se hace a través de línea de comandos con conexión SSH desde un equipo portátil *Apple Macintosh* con Sistema Operativo *Mac OS X*.

Para facilitar la exploración del disco del servidor se establece también una conexión SSHFS (Secure Shell FileSystem), que permita montar remotamente el disco mencionado en el sistema del portátil. Para ellos se utiliza el paquete *MacFuse* [25], que permite utilizar diversos sistemas de archivos sobre *Mac OS X*. Como GUI (Graphic User Interface) para este sistema se cuenta con otro software denominado *MacFusion*[26], que permite eliminar el uso de la línea de comandos para el montaje y exploración del disco montado. Ambos constituyen proyectos de software libre y gratuito.

3.3.1.1 Identificación de Aplicaciones

Además de esta generación de informes, se considera útil para la caracterización de la traza original, la capacidad de identificación de tráfico por aplicaciones. El programa implementado cuenta por tanto con una funcionalidad que permite la identificación de la aplicación que genera cada flujo.

Para esta tarea existen distintas aproximaciones, basadas en las siguiente metodologías:

- **Análisis de firmas características:**

  Durante el intercambio de información entre dos aplicaciones, los primeros paquetes suelen contener información discriminante en forma de cadenas características, o firmas. Estas firmas iniciales, si están previamente identificadas, permiten hacer una asociación fácil entre flujo y aplicación en uso reduciendo, además, el coste computacional que implicaría el análisis de toda la carga útil (o payload) de un flujo, centrándose sólo en el inicio del mismo. Por otro lado, el coste de memoria estará relacionado con la cantidad de datos de ese mismo payload que se almacena para análisis, por lo que, consecuentemente, el programa implementado permite establecer un máximo número de Bytes y paquetes a conservar del payload de cada flujo.

  Las firmas, como se ha comentado, deben estar previamente identificadas y almacenadas en archivos para su comparación. Las firmas utilizadas en esta caso están disponibles en [27] como parte del proyecto L7-filter.

- **Relación Puerto-Aplicación:**

  Aunque este caso se da cada vez menos, existen aplicaciones que suelen utilizar siempre los mismos puertos para realizar sus conexiones. Estas asociaciones, a pesar de estar cambiando en muchos casos hacia conexiones en puertos no comunes, existen todavía para un número limitado de aplicaciones, por lo que su uso se considera aún útil[28].

- **Análisis de protocolo:**

  El conocimiento del funcionamiento de protocolos concretos puede también facilitar la identificación. Para ellos se establecen reglas basadas en características concretas de estos. Como ejemplo, para una transferencia FTP o SIP se realiza una conexión de control y después otra que transmite los datos. Distinguir esto permite hacer una asociación.
  El programa implementa reglas para la identificación para estos, de manera que en cuanto se detecta un flujo de señalización FTP o SIP, se genera una regla dinámica que determina a si existe tráfico de estos protocolos .

3.3.1.2 Establecimiento de Parámetros de Flujo

Como ya se ha descrito, un informe *netflow* incluye, por cada entrada, la quíntupla básica que define un flujo de red, esto es, *IP origen, IP destino, Puerto origen, Puerto destino* y *Protocolo*.

La definición de un flujo pasa también por el establecimiento de un tiempo máximo entre llegadas de paquetes. Para elegir los tiempos a aplicar se acudió a los informes de otros grupos de investigación y a los tiempos de referencia que consideran adecuados para poder seguir la tendencia actual y obtener resultados más contrastables con otros estudios anteriores o contemporáneos.

Estudios anteriormente mencionados, como [10], utilizan un tiempo máximo de 120 segundos, basándose en la referencia dada por [29] y [30].

Otros, como [21] usanlos 30 segundos empleados por la red Suiza SWITCH, de donde obtienen sus trazas de datos. [31] usa también este tiempo, referenciando [32], que es a la vez de los mismo autores, pero que no parece a su vez dar una justificación clara de su decisión.

Existe otro intervalo, alrededor de los 60 o 64 segundos, que se menciona en [33], con 60 segundos, y en [17], con 64. [15] estudia varios casos y obtieneresultados en función de este tiempo, pero cuando usa un tiempo fijo, emplea también este intervalo.

Por último, y al tratarse de uno de los grandes fabricantes de equipos con capacidad de monitorización mediante *netflows*, se observa que Cisco utiliza como intervalo estándar entre llegadas 15 segundos [34].

Se considera esencial adaptarse al estándar *de facto* de la industria, para obtener resultados contrastables, por lo que se elige un tiempo entre llegadas máximo de 15 segundos.

Entre los demás tiempos barajados, y para no utilizar un rango demasiado cercano a los 15 ya adoptados, se opta también por descartar los 30 y 60 a favor de 120, que probablemente podrá aportar resultados más diferenciados y por lo tanto, probablemente más interesantes a la hora de su comparativa.

A la hora de adoptar un tipo de muestreo, se consideró, teniendo en cuenta el estudio del estado del arte, que el empleo de muestreo por paquetes de manera determinista o aleatoria se podía considerar equivalente en lo que respecta a la bondad de sus estadísticas. Se realizarán los experimentos en ambos casos, aunque a la hora de abordar un análisis y proceso de las estadísticas obtenidas, el estudio se centrará en el caso determinista.

Las tasas escogidas para las distintas simulaciones serán de 100%, 50%, 25%, 10% y 1%.

Para el caso de muestreo aleatorio, la correspondencia es directa, de manera que se genera un número aleatorio a cada llegada entre 1 y 100, y se establece el umbral correspondiente, determinando la captura o rechazo de cada paquete.

Para el determinista, se realiza una captura de cada k llegadas, de manera que k permita alcanzar la tasa de muestreo deseada. Dada una tasa X, en tanto por ciento, el valor de k correspondiente será:

$$100 \cdot \left(\frac{1}{X} \right)$$

En las simulaciones deterministas se establece también otro parámetro, que es el *offset*. Este *offset* establece el número de paquete por el que se empieza a muestrear desde el inicio de traza, de modo que no es necesario comenzar muestreando siempre el primero. Esta característica se considera que podría resultar útil a la hora de agregar datos de

distintas simulaciones con la misma tasa, ya que podría permitir una mayor cobertura del número de paquetes total.

Este uso del *offset, como se muestra en la Figura 3-1,* permite simular una situación en la que dos o más puntos de recolección de una red realizan un muestreo a la misma tasa, pero con un desfase que garantiza que el mismo paquete no sea tenido en cuenta dos veces, evitando el problema de la duplicidad de los datos obtenidos, y aumentando la cobertura sin necesidad de usar una tasa mayor.

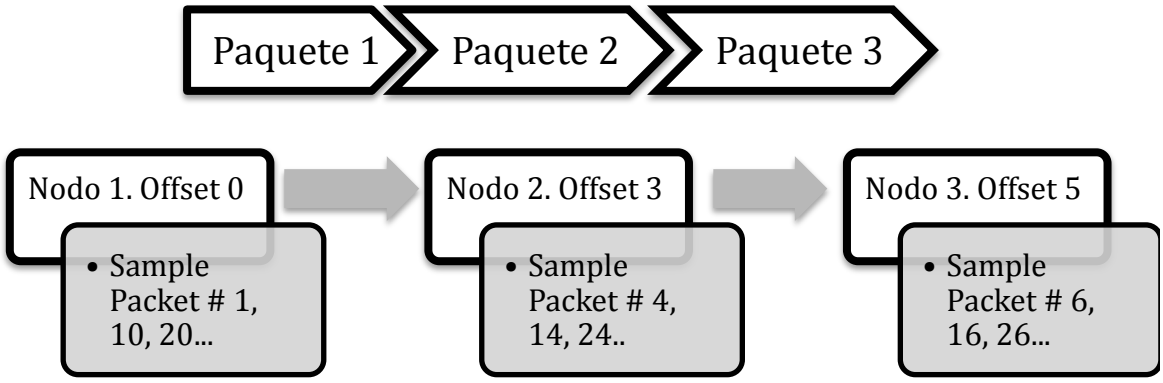


Ilustración 3-1: Muestreo de paquetes de traza según valores de offset

La agregación de los informes obtenidos con este sistema y la comparación de sus resultados con aquellos obtenidos del análisis de la traza original constituirá uno de los puntos clave del presente estudio.

3.3.1.3 Esquema de funcionamiento

A continuación se muestra cuál es el funcionamiento secuencial del programa generador de flujos:

1. Inicio. Evaluación de parámetros de ejecución. Creación de ficheros de salida y tablas hash. Carga de firmas[27] y lista de puertos[35].

2. Según tipo de muestreo, aplicación de discriminador para lectura de cada paquete. Para determinista se usa un contador, tomando un paquete cada k, para aleatorio se aplica un umbral a un número acotado generado al azar. Si el paquete se evalúa, se pasa el siguiente punto.

3. Extracción de carga útil. Se comprueba que el paquete es TCP/UDP, si no se descarta. Si el paquete no contiene carga útil este es también descartado.

4. Identificación del flujo del paquete y actualización de la tabla de flujos. Se comprueba también la existencia de flujos que hayan caducado por el tiempo máximo entre llegadas en la tabla de flujos activos. Se identifica cada flujo de manera única utilizando una función equivalente a la presentada en [16]. En el caso de existir previamente y de no contener ya el número máximo de paquetes, se actualizan los valores de la entrada correspondiente, en caso negativo, se crea una nueva entrada.

5. Aplicación de técnicas identificación de aplicación. Se ejecutan en orden, primero por firma, después por puertos y por último se aplican reglas de protocolo. El nombre de las aplicaciones identificadas se guarda en la tabla para el flujo correspondiente.

6. Por cada flujo inactivo se calculan sus métricas y se exporta al fichero de estadísticas. En el caso de finalizar la traza, se calculan también las métricas del resto de flujos aún activos y se genera una estadística global de porcentaje de aplicaciones.

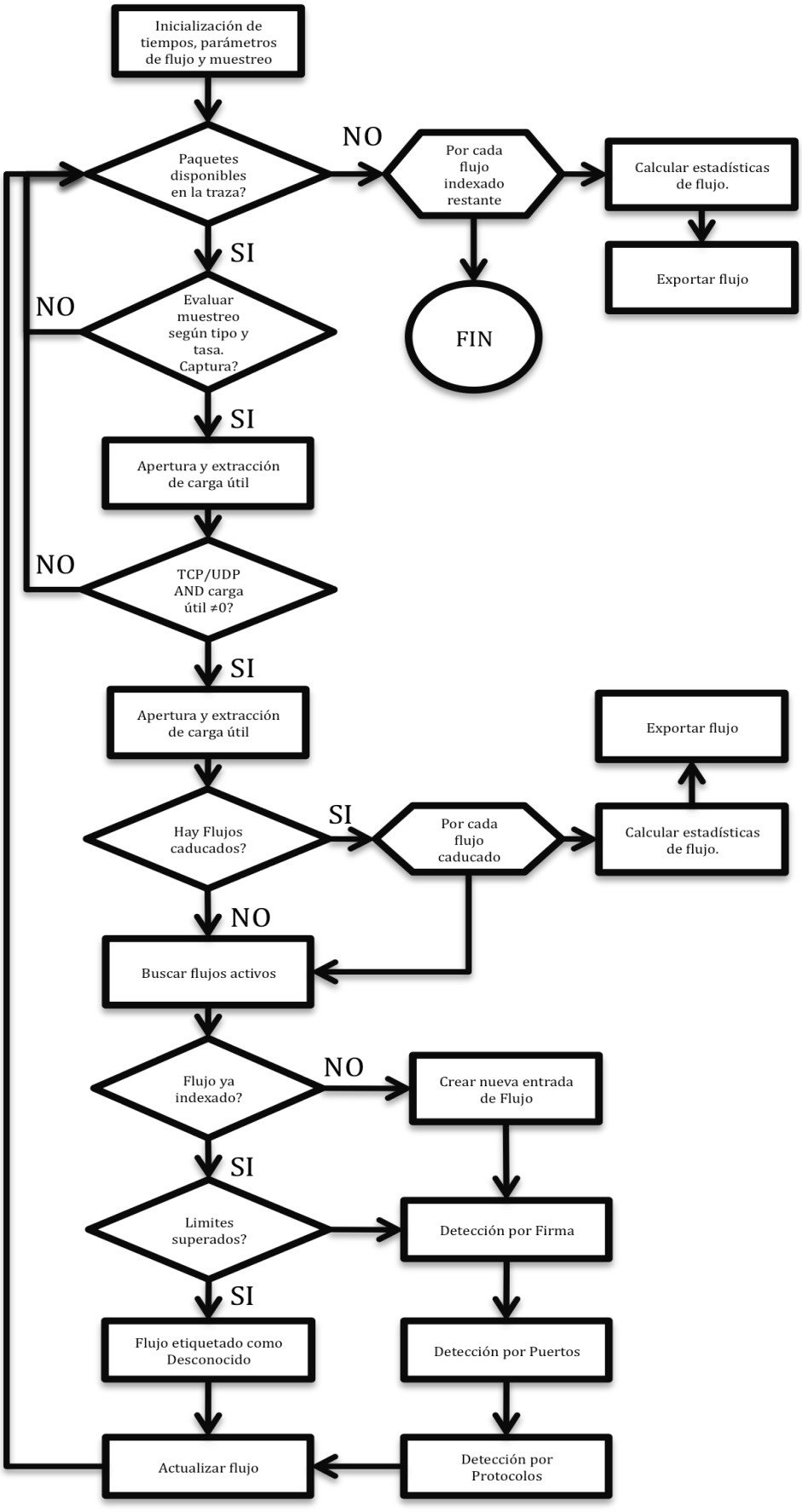


Ilustración 3-2: Funcionamiento de generador de flujos con identificación de aplicaciones

3.3.1.4 Listado de pruebas para muestreo simple

Para lograr una cobertura total de la traza de datos para cada una de las tasas elegidas, se realizan *k* simulaciones por cada una, con offsets que van de 0 a *k-1*. De este modo, la simulaciones que se realizarán, para el caso de muestreo simple, tendrán los siguientes parámetros:

- Aleatorio:
  - Interarrival: 120 segundos
    - Tasa 100%: 1 ejecución
    - Tasa 50%: 2 ejecuciones
    - Tasa 25%: 4 ejecuciones
    - Tasa 10%: 10 ejecuciones
    - Tasa 1%: 100 ejecuciones
  - Interarrival: 15 segundos
    - Tasa 100%: 1 ejecución
    - Tasa 50%: 2 ejecuciones
    - Tasa 25%: 4 ejecuciones
    - Tasa 10%: 10 ejecuciones
    - Tasa 1%: 100 ejecuciones
- Determinista:
  - Interarrival: 120 segundos
    - Tasa 100%. k-1. Offset- 0. 1 ejecución.
    - Tasa 50%: k-2. Offset 0-1. 2 ejecuciones.
    - Tasa 25%: k-4. Offset 0-3. 4 ejecuciones.
    - Tasa 10%: k-10. Offset 0-9. 10 ejecuciones.
    - Tasa 1%: k-100. Offset 1-99. 100 ejecuciones.
  - Interarrival: 15 segundos
    - Tasa 100%. k-1. Offset- 0. 1 ejecuciones.
    - Tasa 50%: k-2. Offset 0-1. 2 ejecuciones.
    - Tasa 25%: k-4. Offset 0-3. 4 ejecuciones.
    - Tasa 10%: k-10. Offset 0-9. 10 ejecuciones.
    - Tasa 1%: k-100. Offset 1-99. 100 ejecuciones.

Esto hace un total de 468 simulaciones a realizar. Al tratarse la traza original de una captura de más de 27 Gigabytes, el proceso se antoja largo y muy demandante en cuanto a ocupación del espacio en disco, ya que la exportación de las estadísticas genera también ficheros de gran tamaño.
Estas estadísticas elaboradas por el generador ofrecen una serie de datos por cada flujo detectado siguiendo las restricciones anteriormente establecidas mediante los parámetros de tiempo entre llegadas y tasa de muestreo.

A continuación se muestran los campos de estos informes de salida que se emplean en este estudio:

- **Inicio**:
  Tiempo de inicio de la flujo detectado. Es un timestamp en formato Unix (número de segundos transcurridos desde el 1 de Enero de 1970) que marca el tiempo de llegada del primer paquete con una 5-tupla que no existía previamente en la tabla.

- **Fin**:
  Tiempo de finalización de flujo detectado. Timestamp en formato Unix que marca el tiempo de llegada del último paquete incluido en la estadística del flujo antes de su exportación. Esta exportación se puede dar por sobrepaso del tiempo máximo entre llegadas o del tiempo máximo de flujo.

- **Duración**:
  Intervalo de duración del flujo detectado en segundos.

- **Tamaño**:
  El número de bytes total del payload de los paquetes incluidos en el flujo de red.

- **Número de paquetes:**
  Número total de paquetes pertenecientes al flujo detectado. Estos se contabilizan siempre que pertenezcan a los protocolos TCP o UDP.

- **Throughput**

  Este campo determina la tasa media en Megabits por segundo, es decir el ancho de banda media usado, de los datos que se han detectado de un flujo, desde su detección hasta su exportación.

3.3.1.5 Tiempos de ejecución y Tamaño de Informes

Las llamadas al generador se realizan vía *shellscript*, especificando mediante estos los parámetros de cada simulación, que se tratan de realizar de manera automática.

Para poder presentar una estimación de los tiempos de simulación y tamaños de informes exportados, la simulación de la traza original, de 27.6GB, con muestreo al 100% y tiempo entre llegadas máximo de 15 segundos se realiza en 8 horas, y el informe resultante, en formato .csv, ocupa 4.54 Gigabytes. El resto de simulaciones, con tasas menores no siguen una tendencia linealmente decreciente en cuanto al tamaño total de los informes exportados o el tiempo de ejecución, tendiendo estos a ocupar mayor espacio relativo. De este modo, por ejemplo, si se quisiera abarcar toda la traza de datos agregando los 100 informes exportados de las ejecuciones con muestreo determinista al 1% para el mismo tiempo entre llegadas, correspondientes a los 100 valores de *offset* distintos posibles, estos ocupanen total 11.77 Gigabytes, más del doble que el informe al 100%. La tendencia relativa de todas las tasas empleadas se puede observar en la figura 3-3.

Esta es una característica que, si bien esperada, se considera interesante mencionar, ya que se prueba cómo, en efecto, un muestreo a una tasa un tanto por ciento inferior a otra no permite necesariamente ahorrar ese equivalente en espacio de disco para informes. Esta característica hace que el empleo de sistemas *netflow* muestreados pueda llegar a generar a la larga un requerimiento excesivo en cuanto a almacenamiento de datos de informes, si se pretende realizar una cobertura relativamente completa de los datos que atraviesan la red.

Se muestra a continuación una gráfica comparativa que pone de manifiesto este fenómeno, para el caso de 15 segundos como tiempo máximo entre llegadas. Por cada tasa se muestra el tamaño del informe realizado con *offset* 0 y a la vez, la suma de todos los demás informes realizados con los demás valores de *offset* distintos necesarios para almacenar los datos de toda la traza completa. De nuevo, a modo de ejemplo, para una tasa del 10%, hay 10 offsets distintos a aplicar para cubrir toda la traza, y por lo tanto 10 informes.

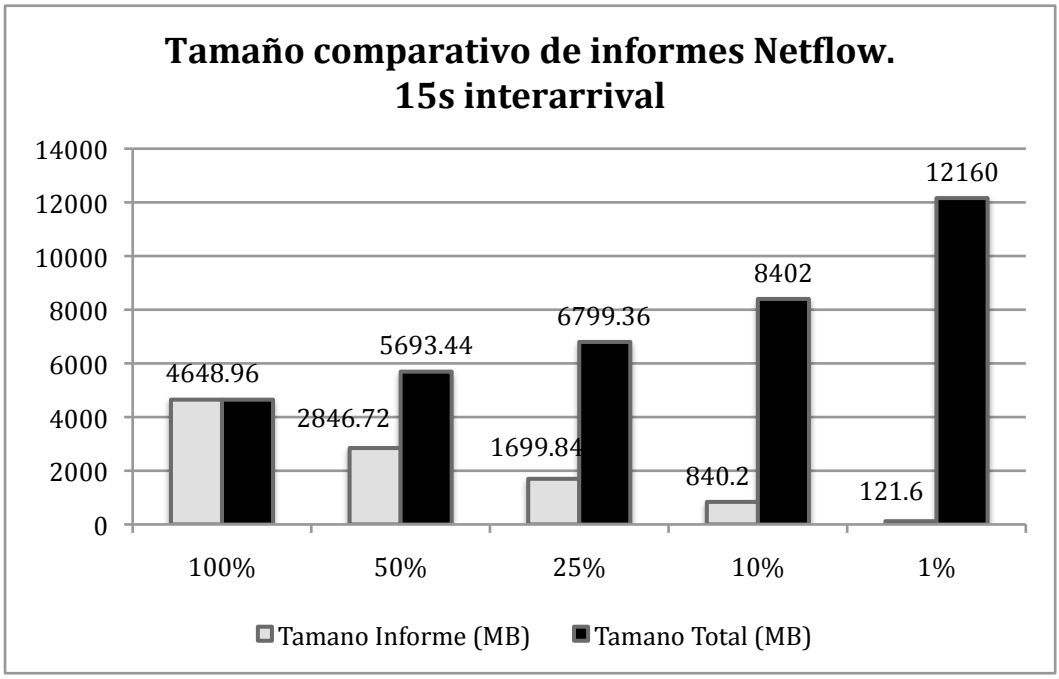


Ilustración 3-3: Tamaños relativos de informes agregados a diferentes tasas de muestreo

Queda por otro lado patente, a nivel de ejecución de pruebas para el estudio, que la realización de las mismas requiere tiempos relativamente largos y un espacio disponible en disco más que considerable. A este efecto se hizo necesario en momentos del proceso el volcado de datos a otras unidades y el montaje en el servidor de discos nuevos que permitieran el almacenamiento de los datos.

3.3.2 SW Matlab

Después de obtener las estadísticas exportadas por el generador, es necesario el empleo de otra herramienta que permita hacer un análisis de los datos obtenidos y extraer métricas, no sólo directamente de los informes, sino mediante otros cálculos que ofrezcan una visión más completa de la naturaleza del tráfico muestreado.

Se hace vital, por otro lado, la capacidad de cálculo de otras funciones más complejas para la presentación de resultados. Para ello se emplean gráficas, como histogramas o funciones de distribución.

La comparativa entre resultados se realiza de manera, tanto visual, a través de la comparación de las gráficas obtenidas para las métricas escogidas, como mediante otros tests que permitan evaluar cuantitativamente las diferencias entre las mismas.

Se elige una versión del entorno de simulación *Matlab* como herramienta para este cometido. Se instalará sobre un equipo portátil *Apple Macintosh* con sistema operativo *Mac OS X*. El equipo cuenta con un procesador *Intel Core 2 Duo* a 2,4GHz y una memoria SDRAM a 667MHz de 2 Gigabytes.

Para la extracción de datos se comprueba cómo el gran tamaño de los informes .csv hace imposible su tratamiento de manera directa, que provoca un desplome del sistema por falta de memoria, por lo que se establece como paso previo al procesado de datos una extracción de cada una de las columnas con los campos ya mencionados en archivos de texto independientes.

Esta extracción de columnas se antoja de nuevo un proceso largo, ya sea cuando el acceso a los informes se realiza de manera remota, como si se hace una transferencia previa de los datos a un disco local que, de nuevo por problemas de espacio, debe ser externo con comunicación USB 2.0. Esta eventualidad retrasa el análisis y provoca un mayor requerimiento de espacio en disco, al realizarse en esencia una duplicación de los datos. El factor positivo es que, una vez extraídas las columnas de datos, *Matlab* es capaz de guardarlas con rapidez en memoria como matrices independientes con una sola llamada a la función *load*.

Una vez lograda la extracción, se implementan una serie de funciones y *scripts* para las siguientes tareas:

- Identificación de métricas de interés.
- Elaboración de Histogramas.
- Cálculo de Funciones de Distribución.
- Ejecución de tests de bondad de ajuste.

- Agregación de Estadísticas para muestreo distribuido.

- Creación y presentación de Gráficos.

3.4 Características Medidas

Se considera necesario para este estudio el adquirir conocimiento sobre la naturaleza del tráfico a estudio mediante la traza obtenida.

Este análisis de tráfico se realiza mediante la obtención y observación de un número de métricas, a través de las cuales se pretenden obtener características sobre el mismo.

Estas métricas serán útiles, no sólo para el simple conocimiento de la dinámica del tráfico, sino también para poder utilizar algunas de estas como "unidad de medida" a la hora de valorar el impacto que el muestreo realizado con sus distintos parámetros puede tener sobre las estadísticas obtenidas.

La función distribución de probabilidadde estas métricas para distintos muestreos, ya sea este simple o distribuido, se utilizará para evaluar la bondad de ajuste. Esto puede aportar una valoración objetiva sobre la diferenciación cuantitativa que pueden llegar a tener los informes *netflow* obtenidos con distintos métodos y parámetros con respecto a los de la traza original muestreada al 100%.

Se muestran a continuación las métricas consideradas de interés para el estudio.

- Duración de Flujo, medido en segundos.

- Número de Paquetes por Flujo.

- Numero de Bytes por Flujo, o Tamaño de Flujo.

- Throughput, o caudal empleado por Flujo.

- Número total de flujos detectados.

- Ranking de Flujos, esto es, identificación de los 100 flujos de mayor tamaño en orden.

- Bytes por IP. Con esta medida se trata de obtener la distribución del tráfico saliente en cuanto a las IPs destino de cada flujo.

- Flujos solteros. Aquellos que no tienen correspondencia en otro sentido. De algún modo se le podría denominar *sesiones unidireccionales*.

- Flash Flows. Ya se hizo referencia a los mismos y al posible interés de su estudio en el estado del arte. Son aquellos que duran menos de un segundo, contienen menos de 3 paquetes y suman un número total de Bytes menor que 500.

  Porcentaje de tráfico total según aplicaciones. Esta métrica pretende dar una visión algo más específica sobre la naturaleza del tráfico, no estrictamente relacionada con las definiciones de flujo. Se pretende con esta métrica, además, aprovechar una funcionalidad muy útil con la que cuenta nuestro generador.

3.5 Resultados

A continuación se muestran los resultados obtenidos tras la extracción de informes *netflow* para la traza original con un muestreo al 100% y con dos tiempos máximos entre llegadas para definir flujos.

3.5.1 Duración

Las siguientes gráficas corresponden a funciones de distribución de probabilidad acumulada (cdf) de la duración de los flujos, en segundos, para los dos tiempos entre llegadas de paquetes establecidos (esto es 15 y 120 segundos). El eje de las abscisas se ha dispuesto en unidades logarítmicas, puesto que las mayores variaciones se dan en los primeros rangos, siendo este en total relativamente extenso. El eje de ordenadas se muestra en coordenadas naturales.

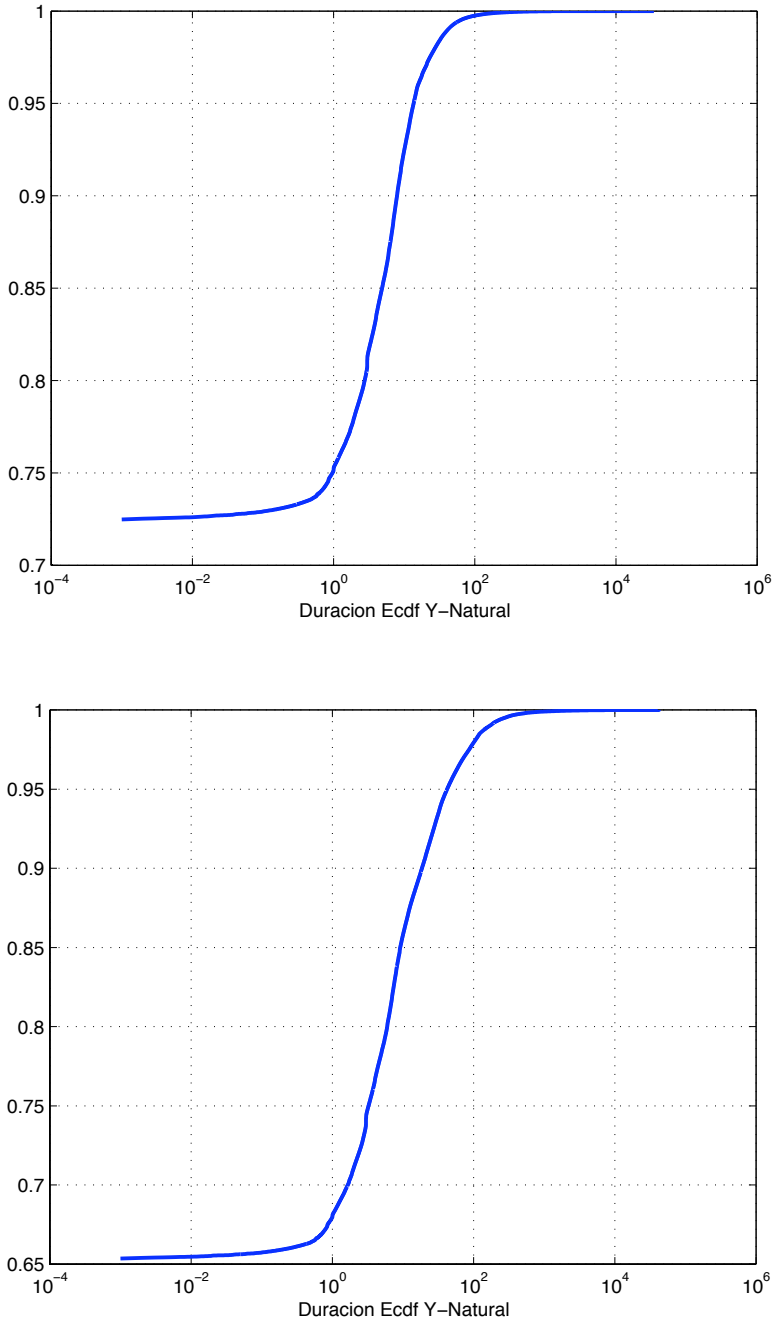


Ilustración 3-4: CDF Duración Traza original

En ambos casos se observa cómo una gran mayoría del total de los flujos registrados dura menos de un segundo, muchos de ellos como veremos más adelante, son flujos de tan sólo un paquete y por tanto duración nula. En concreto, en el primer caso estos flujos representan más del 70% y en torno al 65% en el segundo caso.

En cuanto a flujos mayores de 1 segundo podemos comprobar que típicamente la duración de los flujos de red están el rango de 1-100 segundos distribuidos de forma cuasi-uniforme

tras haber aplicado un transformación logarítmica al eje x. Es muy infrecuente encontrar flujos de duración mayor a 100 segundos (en torno a un 2% en el segundo caso) y la probabilidad de flujos mayores a 1000 segundos es aun menor pero es importante destacar que no es nula.En concreto las duraciones máximas de los flujos de acuerdo a su interarrival son:

- Interarrival 120 segundos. Flujo máximo = 42208 segundos. 11,72 horas.

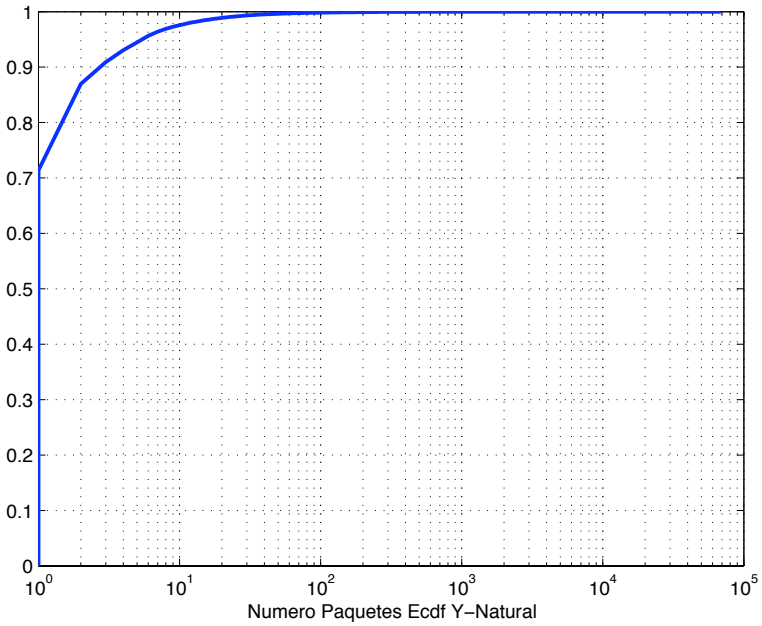
- Interarrival 15 segundos. Flujo máximo = 34595 segundos. 9,6 horas.

Esta amplia diferencia entre el valor medio y el máximo de distribución es característico de aquellas distribuciones denominadas de *cola pesada* [36], y se podrá comprobar cómo resulta una constante en la mayoría de las métricas obtenidas. Este fenómeno es mencionado con más detalle en la sección "*Impacto del Muestreo*".

3.5.2 Número de Paquetes.

Se muestran las cdfs del número de paquetes por flujo.

De nuevo, el eje de ordenadasse muestra en coordenadas logarítmicas, mientras que el eje de abscisas se muestra en coordenadas naturales.



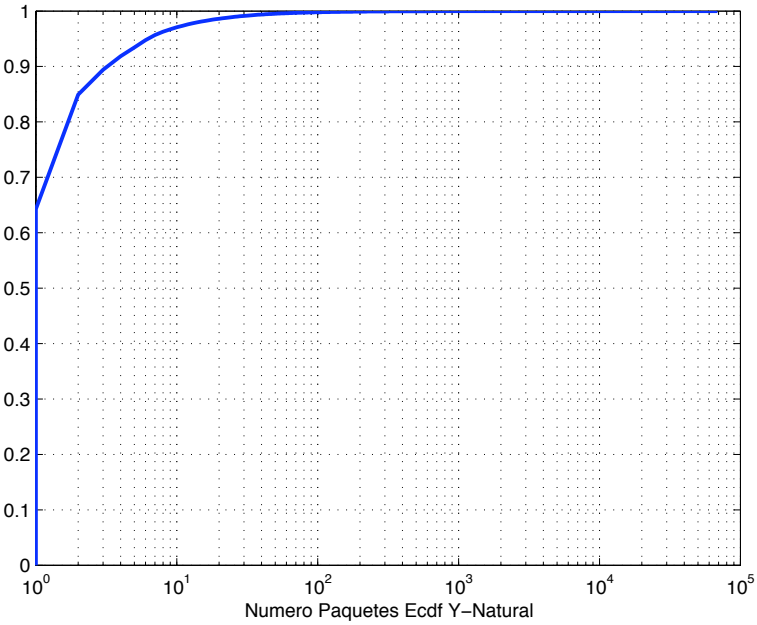


Ilustración 3-5: CDF Número de Paquetes Traza Original

Se observa cómo la mayor concentración se encuentra en torno al valor mínimo, es decir, un paquete por flujo, lo cual es coherente con los resultados mostrados previamentePara 15 segundos aparecen más del 71.3% de flujos con un único paquete, mientras que para 120 segundos esta concentración es del 64.2% aproximadamente. La anticipación de este fenómeno no deja de ser obvia, ya que un menor máximo tiempo entre paquetes genera nuevas entradas.

La distribución sigue un esquema de cola pesada que se aproxima lentamente hacia el valor máximo, que tanto para el caso de 120 segundos como para 15, se alcanza en 68812 paquetes.

3.5.3 Número de Bytes.

El número de bytes por flujo a nivel de aplicación es una métrica que complementa al número de paquetes. Esta medida acumula el tamaño en bytes del payload de cada paquete perteneciente a cada flujo.De este modo, habiendo implementado el generador de flujos del tal modo que se ignoran los paquetes sin carga útil, el tamaño mínimo de un flujo es 1.

Se muestran de nuevo las distribuciones de la misma para 15 y 120 segundos de tiempo máximo entre llegadas, respectivamente.

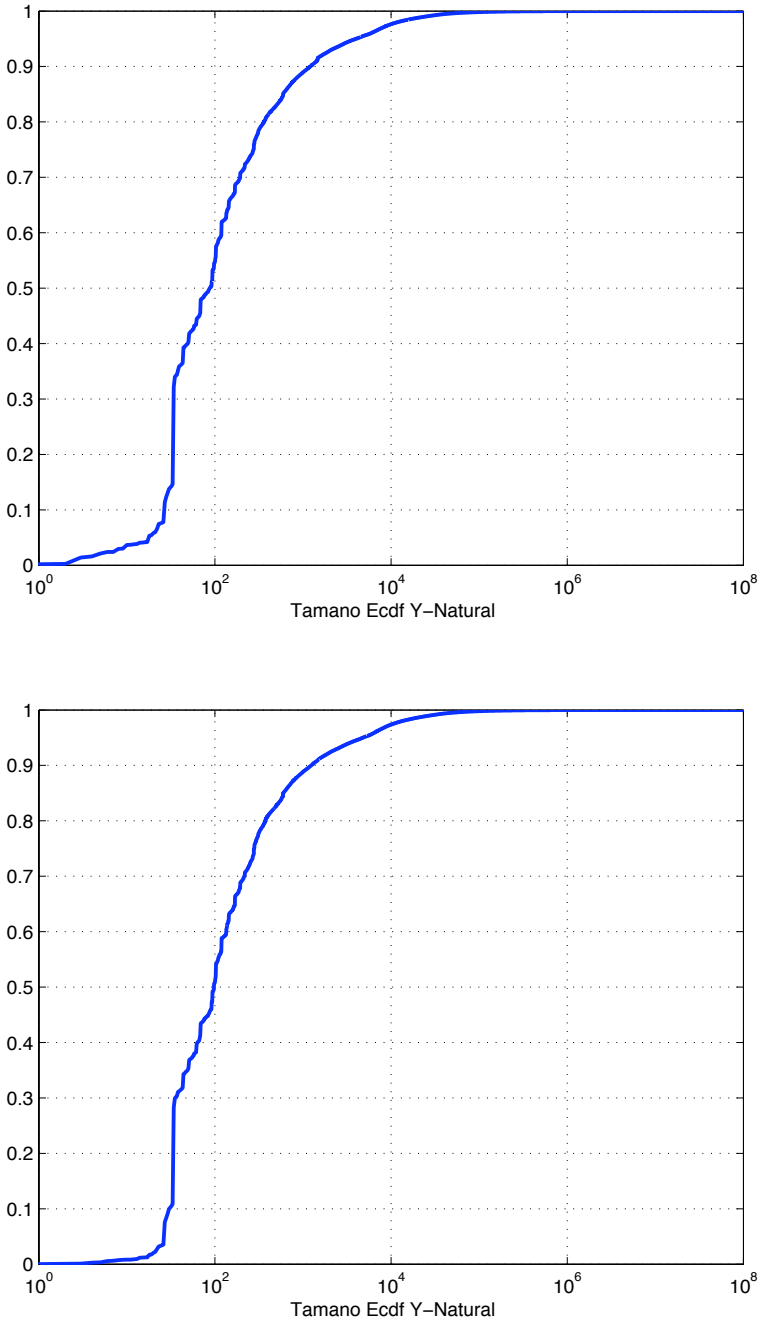


Ilustración 3-6: CDF Número de Bytes Traza Original

El contraste visual entre ambas distribuciones hace patente la poca trascendencia que tiene en esta métrica el uso de distintos tiempos entre llegadas, tan sólo leves diferencias pueden ser advertidas en la cola.

Se observa una moda en torno a los 34 bytes, para ambos casos. La existencia de esta moda, y los resultados obtenidos según análisis de las aplicaciones identificadas e

inspección manual de los paquetes correspondientes apuntan a una correspondencia con peticiones DNS a páginas como *hotmail.com*, *tuenti.com* o *time.microsoft.com.*

3.5.4 Throughput.

El throughput, en Mbps, o caudal de datos por cada flujo detectado, es una métrica que se extrae de manera sencilla a partir del tamaño y duración. Aunque no proporciona en esencia datos nuevos sobre las características del tráfico, se considera útil poder observar su distribución. Se ha definido el throughput de un flujo de duración nula como cero:

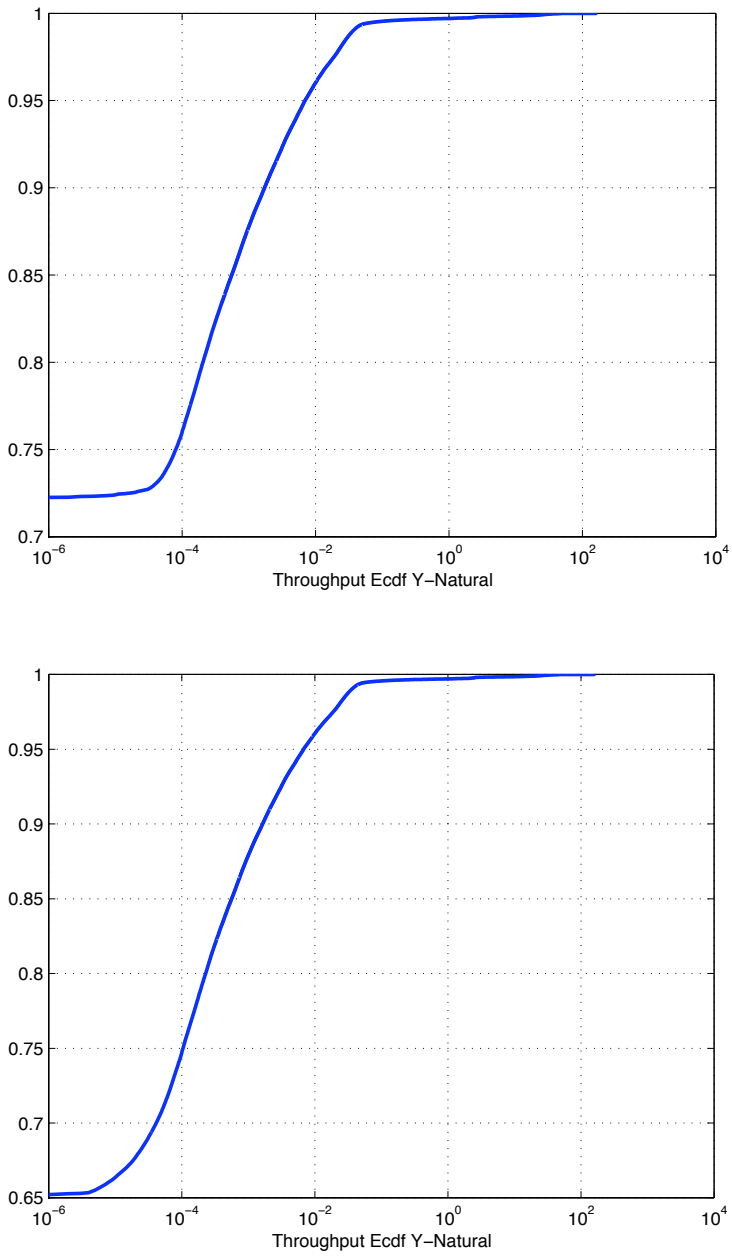


Ilustración 3-7: CDF Número de Bytes Traza Original

De nuevo existe una tendencia en el caso de 15 segundos de interarrival máximo a mostrar una distribución con mayor concentración al inicio de la gráfica, es decir, con mínimo ancho de banda ocupado. En cuanto a la diferencia del valor máximo entre ambas simulaciones, la tendencia resulta igual que para la métrica de tamaño, donde este valor resultaba el mismo para ambos. En este caso, este máximo, si bien no es exactamente igual, es muy parecido.

En el caso de 15 segundos, el 72.23% de flujos tiene un valor de caudal nulo, que corresponde a flujos de duración0, es decir, sólo abarcan un paquete, mientras que el máximo ocupado por uno de ellos es de 166,74 Mbps, medido en un flujo de apenas unos pocos paquetes y, por tanto, poco representativo de la verdadera capacidad de la línea.

Para 120 segundos el 65,15% de los flujos detectados tiene un caudal también nulo, aproximadamente un 7% menos, y el valor máximo alcanza los 166,42 Mbps, un resultado que se puede considerar virtualmente idéntico que para el caso anterior.

Por último, cabe notar que la distribución muestra un crecimiento casi perfectamente logarítmico en un rango apreciable de la función para ambos casos.

3.5.5 Número total de flujos

Esta métrica proporciona poca información de momento a la hora de caracterizar la traza original, pero será un dato de referencia útil para compararlo con los resultados según distintas tasas de muestreo.

- Para 15 segundos entre llegadas, se detectan 10,671,109 flujos.

- Para 120 segundos entre llegadas, se detectan 12,647,520 flujos, aproximadamente un 18.5% más que para el caso anterior.

3.5.6 Ranking de Flujos.

En esta sección se muestra una lista con los 100 flujos más significativos en términos de su tamaño.

P destino	P origen	Protocolo	Payload Bytes
80	2746	6	10278926
80	2337	6	10390098
80	49309	6	10390098
11307	1137	6	10396561
27150	39749	6	10500754
1935	4989	6	10532357
80	50378	6	10554617
80	1402	6	10600208
110	1076	6	10692666
5958	1076	6	10765420
80	2071	6	10885760
80	3421	6	11110552
80	1085	6	11110552
2082	37392	6	11258740
47599	1384	6	11303404
37805	55716	6	11316759
11949	20657	6	11316759
27150	39749	6	11341280
80	2746	6	11403261
24160	3183	6	11429117
80	1094	6	11469006
2584	2148	6	11469006
3803	4583	6	11646336
2528	20657	6	11732278
46632	2446	6	11742192
3467	15162	6	12062418
2082	37392	6	12064163
24721	32681	6	12137192
19914	18012	6	12187228
59018	20657	6	12190383
1277	2148	6	12334440
80	1295	6	12334440
80	1304	6	12731163
18120	61241	6	12731163
554	1323	6	13176748
80	1410	6	13539372
46632	1557	6	13791694
80	1992	6	14012704
14806	2885	6	14300760
80	1114	6	14611680
43632	2397	6	14706580
80	49556	6	14712420
110	1051	6	14712420

23948	29677	6	14772280
80	1194	6	14850729
26204	18012	6	14905140
80	49556	6	15085823
80	1821	6	15368232
80	1071	6	15368232
554	1110	6	15388400
80	49556	6	15707273
80	1821	6	15897145
24137	29677	6	16773963
1501	39749	6	16773963
80	1567	6	16845076
80	1104	6	17177977
80	1686	6	17786269
80	1962	6	17833514
80	1510	6	17841429
3921	18012	6	18423139
80	1216	6	18718284
80	1942	6	18932308
58850	1583	6	19278852
24212	29677	6	19278852
8870	8870	17	19772780
80	1230	6	19772780
110	1753	6	20373501
80	1152	6	21250596
5846	1057	6	21250596
80	49385	6	21568926
80	1363	6	21568926
80	1200	6	21611863
80	1664	6	21828404
80	1500	6	22616910
80	1085	6	23156442
24212	29677	6	23329130
80	1518	6	23329130
80	1501	6	24757220
80	1213	6	24757220
80	1686	6	25032469
80	1962	6	27255280
80	1510	6	27255280
80	1216	6	27273461
80	1942	6	28281874
58850	1583	6	31559318
24212	29677	6	31769355
80	1230	6	32522644
110	1753	6	34147207

80	1152	6	35572801
5846	1057	6	37734342
80	49385	6	49991563
80	1363	6	49991563
80	1200	6	59047418
80	1664	6	59238540
80	1500	6	59350460
80	1085	6	70141058
24212	29677	6	100121139
80	1518	6	100121139
80	1501	6	100462777
80	1213	6	100462777

Tabla 3-1: Ranking the Flujos. Traza Completa

Se comprueba como la mayoría de estos grandes flujos corresponden a conexiones HTTP. El primero de todos, tras inspección manual, corresponde a una pagina de descargas directas, *rapidshare.com*, y las dos siguientes no tienen correspondencia a día de hoy al hacer una petición DNS. Siguiendo conexiones al puerto 80 en la lista, la página *rapidshare.com* parece ser la más común. Otros flujos, con puertos no conocidos, y vista la gran cantidad de datos que contienen, podrían deberse también a servicios de intercambio de archivos. Entre estos, aparece en noveno lugar una conexión al puerto 110, que corresponde con la conexión a un servidor de correo.

De estos resultados se podría extraer que los flujos de mayor tamaño corresponden, en efecto, a descargas directas a través de páginas Web y, probablemente, a aplicaciones P2P, con alguna excepción, como el acceso a servidores de correo electrónico.

3.5.7 Bytes por IP.

Esta métrica trata de mostrar cual es la distribución del tráfico de salida de la traza analizada. Con esto se pretende identificar la existencia de rangos o direcciones IP concretas que sean objeto frecuente de conexión. Para 15 segundos entre llegadas, la distribución de este tráfico de salida es la siguiente:

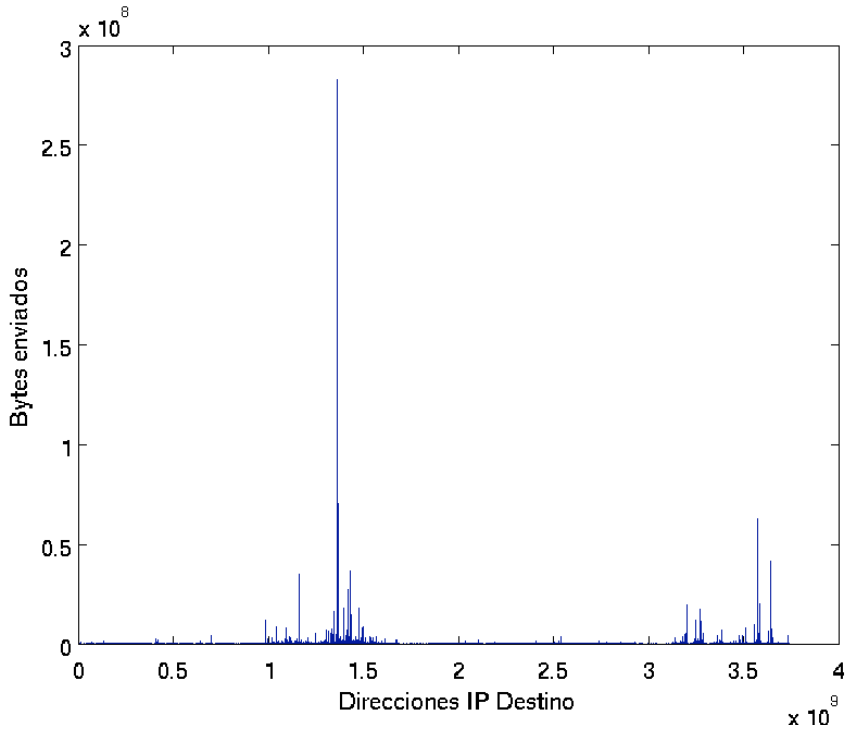


Ilustración 3-8: Bytes por IP. 15s interarrival

Para 120 segundos entre llegadas, la distribución de este tráfico de salida es la siguiente:

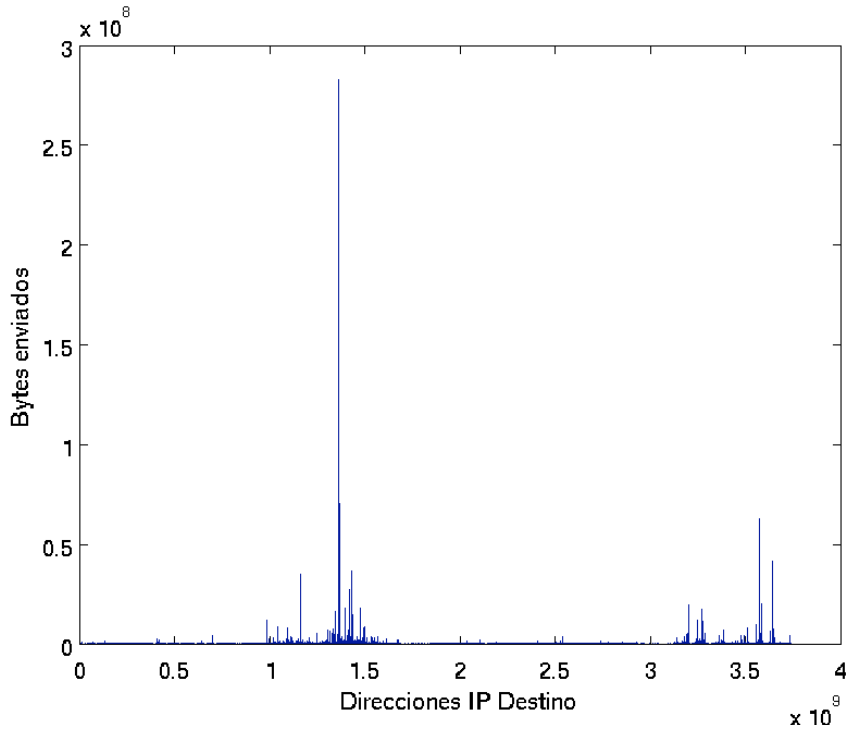


Ilustración 3-9: Bytes por IP. 120s interarrival

Se puede observar cómo la mayor parte del tráfico tiende claramente a acumularse entorno a una dirección, aunque existen otras 5 direcciones que también reciben una cantidad relativamente alta.

3.5.8 Flash Flows.

Se ha explicado en puntos anteriores del documento la razón del estudio de esta métrica, que permite confirmar el fenómeno de la existencia de flujos extremadamente cortos en número de paquetes, tamaño y tiempo de duración:

- Menos de 3 paquetes.
- Menos de 1 segundo de duración.
- Menos de 500 Bytes.

La observación de los mismos puede aportar datos en la identificación de anomalías, y se intuye que los estudios de posibles nuevas definiciones de flujo a adoptar en el futuro puedan hacer uso de la misma para lograr una distribución más uniforme en los informes obtenidos.

En ambos casos de tiempo entre llegadas se ha implementado una búsqueda de este tipo de flujos.

Para 15 segundos de interarrival, de un total de 10671109 flujos detectados, 6732697 de los mismos son Flash Flows. Esto constituye un 63% del total, lo que confirma el gran peso que tienen los mismos en los informes resultantes.

Para 120 segundos de interarrival, los flujos totales detectados son 12647520, siendo 8777569 de estos Flash Flows, es decir un 69%. El porcentaje de los mismos, por lo tanto, es mayor para este caso.

Resulta interesante, por otro lado, que esta concentración de flujos tan grande no se corresponde en absoluto con una gran cantidad de tráfico en cuanto al número de Bytes total. A este efecto, se muestran los datos obtenidos.

Tanto para 15 como para 120 segundos, el número de Bytes total es de 24116720000. El hecho de obtener la misma cantidad para ambos proporciona un cierto grado de validación de la bondad del generador implementado. El número de Bytes correspondientes a flash flows es de 615159576 para 15 segundos, que equivalen a un 2,5508% del total, y para 120 segundos este valor es 760049018 Bytes, es decir, un 3,1515% del total.

Estos resultados justifican la existencia de estudios que tratan de eliminar estos flujos de sus informes, como [37]. Si el sistema de monitorización constituye principalmente un método de tarificación por ancho de banda consumido, la contabilización de estos flujos cortos resulta muy demandante en términos de espacio en disco ocupado por los informes, y sin embargo corresponde a una cantidad de tráfico muy pequeña. Eliminar su análisis en estos casos puede resultar positivo para el objetivo especificado.

3.5.9 Porcentaje por aplicaciones.

La funcionalidad de identificación de aplicaciones del generador de flujos permite la realización de tablas de distribución del tráfico total de la traza analizada. Los resultados para ambos tiempos entre llegadas son los siguientes:

15 segundos interarrival			120 segundos interarrival		
Application	**Bytes**	**Percentage**	**Application**	**Bytes**	**Percentage**
http	13700788335	56.81	http	14662446035	60.8
UnKnown	7610535166	31.56	UnKnown	6202920127	25.72
ssl	874819578	3.63	ssl	984888519	4.08
edonkey	723382370	3	edonkey	967927161	4.01
pop3	287508077	1.19	pop3	287508077	1.19
msnmessenger	174944896	0.73	bittorrent	263072688	1.09
smtp	171424174	0.71	msnmessenger	174738057	0.72
dns	154436997	0.64	smtp	171424917	0.71
bittorrent	127210869	0.53	dns	154434524	0.64
rtsp	90541071	0.38	rtsp	90549443	0.38
skypetoskype	36867632	0.15	freenet	36859111	0.15
nbns	29771495	0.12	rtp	22319816	0.09
stun	21925814	0.09	skypetoskype	15532118	0.06
rtp	21281686	0.09	xunlei	15079611	0.06
xunlei	15415452	0.06	nbns	8404337	0.03
socks	13938820	0.06	smb	8127059	0.03
pplive	13000282	0.05	fasttrack	7540200	0.03
fasttrack	9492047	0.04	socks	7321819	0.03
freenet	7821908	0.03	stun	6718946	0.03
smb	7647017	0.03	pplive	4960241	0.02

imap	4193152	0.02	imap	4650742	0.02
yahoo	3843214	0.02	yahoo	4067396	0.02
netbios	3764748	0.02	netbios	3284706	0.01
napster	2874639	0.01	qq	1859330	0.01
ftp	1681145	0.01	ftp	1681185	0.01
qq	1634621	0.01	aim	1380678	0.01
armagetron	1321118	0.01	armagetron	1333938	0.01
soulseek	1157287	>0.01	napster	1230317	0.01
aim	870126	>0.01	vnc	1218520	0.01
marca	540005	>0.01	soulseek	907910	>0.01
vnc	455790	>0.01	marca	540005	>0.01
rdp	248275	>0.01	rdp	330413	>0.01
sip	229608	>0.01	sip	275952	>0.01
megaupload	184697	>0.01	youTubeClick	203873	>0.01
elMundo	169700	>0.01	megaupload	184697	>0.01
youTubeClick	168905	>0.01	elMundo	169700	>0.01
gnutella	164340	>0.01	gnutella	159467	>0.01
abc	155993	>0.01	abc	156882	>0.01
megauploadAcceso	71870	>0.01	megauploadAcceso	74792	>0.01
ntp	60051	>0.01	ntp	60051	>0.01
jabber	51372	>0.01	jabber	54469	>0.01
elPais	37982	>0.01	irc	44833	>0.01
x11	34556	>0.01	elPais	37982	>0.01
rapidshareAcceso	26536	>0.01	rapidshareAcceso	26536	>0.01
rlogin	14562	>0.01	x11	14938	>0.01
irc	13847	>0.01	h323	4261	>0.01
h323	4261	>0.01	rapidshare	3619	>0.01
rapidshare	3619	>0.01	lpd	1539	>0.01
lpd	1539	>0.01	shoutcast	1462	>0.01
shoutcast	1462	>0.01	imesh	520	>0.01
imesh	520	>0.01	ventrilo	429	>0.01
ventrilo	429	>0.01	msn-filetransfer	138	>0.01
gopher	258	>0.01	pcanywhere	4	>0.01
msn-filetransfer	138	>0.01	**TOTAL**	**22.46GByte**	**100**
battlefield2142	35	>0.01			
pcanywhere	4	>0.01			
TOTAL	**22.46GByte**	**100**			

Tabla 3-2: Porcentaje de aplicaciones identificadas. Traza completa

Para lograr tener una visión más clara de esta distribución, se muestra la misma mediante una agrupación de las aplicaciones y protocolos identificados según las siguientes categorías:

- WWW: Engloba peticiones DNS, HTTP.
- Intercambio: Incluye los programas de intercambio que usan P2P y otros, como FTP o descargas directas vía Web tipo *Megaupload*.

- Remoto + Seguro: En esta sección están las conexiones SSL, que son claramente mayoritarias, y otros accesos como Rlogin, VNC o sesiones X11.
- Mensajería: Incluye servicio de correo electrónico, SMTP, IMAP, POP3, y otros de mensajeria instantanea y voz por IP.
- Entretenimiento: Está constituida por juegos y accesos a páginas de contenidos de noticias.
- Otros: Tráfico cuya aplicación no ha podido ser determinada.

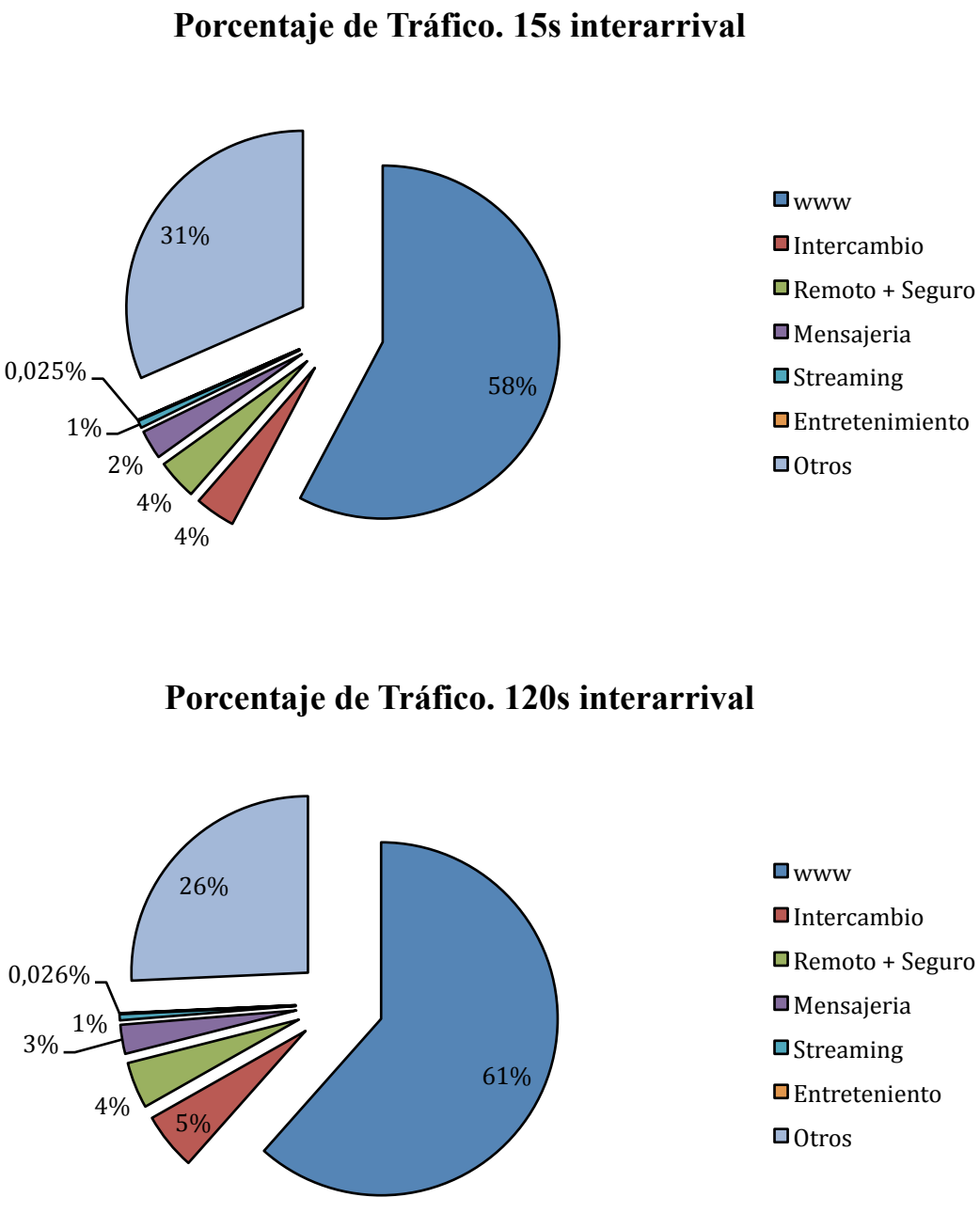


Ilustración 3-10: Porcentaje de tráfico por aplicaciones Traza Original

Se observa cómo, aunque existe una cierta variabilidad, las aplicaciones más utilizadas, como accesos HTTP, SSL, peticiones DNS o el uso de protocolos de correo, como POP3 y SMTP, se sitúan aproximadamente en las mismas posiciones de la tabla. Flujos P2P, debidos a programas como *bittorrent* o *edonkey* tambien son detectados e identificados en posiciones altas, aunque algo más variables. Es interesante notar cómo el tráfico P2P, a pesar de ser significativo, no constituye un volúmen tan grande de tráfico en número de bytes en comparación con el gran número de flujos que genera. Esto es seguramente debido a que este tipo de aplicaciones generan muchas conexiones de control para que los usuarios intercambien y publiciten los recusos disponibles. Estas conexiones apenas transmisten bytes pero generan al menos un flujo por cada una.

Cabe también destacar cómo el porcentaje de flujos con aplicaciones no identificadas es mayor para el caso de 15 segundos que para el de 120. En concreto, la mejora alcanzada con 120 segundos estáentorno al 6%. Esto parece claro, al poder asumir que un mayor intervalo máximo entre llegadas genera flujos más largos, con más paquetes, y por lo tanto, más fáciles de identificar.

4 Impacto del Muestreo

4.1 Introducción

Esta sección muestra el proceso llevado a cabo para comprobar la veracidad de los informes *netflow* obtenidos a través de una técnica de muestreo de paquetes.

El tipo de muestreo adoptado, sobre el que ya se hizo mención en el estado del arte, estará determinado por eventos, donde cada llegada de un paquete genera una apreciación, en la que se decide si analizar el paquete o desecharlo. Esta decisión se tomará de manera determinista, esto es, se tomará un paquete de cada *k* llegadas. Se pudo comprobar en la segunda sección, a través de los resultados de otras investigaciones, cómo este tipo de muestreo se puede considerar equivalente al aleatorio en cuanto a la bondad de sus resultados. Este hecho se ha podido también confirmar mediante la ejecución de pruebas con este tipo de muestreo, con resultados preliminares prácticamente idénticos y que, por tanto, no se incluyen en este apartado.

Este muestreo determinista permite, por otro lado, implementar un sistema que aplique un *offset* determinado a la traza a estudio, de manera que la evaluación de paquetes no comience necesaria y estrictamente por el primero. De este modo, y para una tasa, por ejemplo del 10%, es decir, de 1 paquete analizado por cada 10 llegadas, se pueden aplicar los *offsets* de 0 a 9.

Ejecutando nuestro generador de *netflows* mediante este tipo de captura y con cada uno de los *offsets* permitidos, es posible tomar todos los datos de la traza sin superposición alguna de los mismos. Poder implementar una solución parecida con distintas combinaciones permite simular el funcionamiento de un sistema que hace un muestreo de manera distribuida. La fusión de los informes obtenidos para cada métrica considerada pretende obtener mejores resultados que aquellos que se obtengan mediante muestreo convencional.

Uno de los objetivos, por lo tanto, de este apartado, es hacer una apreciación de los resultados obtenidos en los informes *netflow* mediante muestreo determinista convencional como paso previo al análisis de los que se podrían obtener a través de la agregación de estadísticos obtenidos mediante muestreo distribuido.

En el caso de muestreo convencional, se evaluará el impacto que el uso de distintas tasas y tiempo máximos entre llegadas tiene sobre distintas métricas a considerar. Esto aportará información objetiva y clara sobre qué medidas se ven afectadas y cuánto por el uso de esta técnica.

La sección se divide de manera natural en dos subsecciones principales. La primera subsección se centra en los resultados obtenidos mediante muestreo convencional. En ella se hace primero una descripción del operador que se utilizará para medir el grado de verosimilitud que pueden tener los resultados obtenidos con muestreo frente a los de la traza original sin muestrear, y en la segunda se pasa a continuación a la evaluación de los resultados obtenidos.

En la segunda sección, y para cada métrica a considerar, se efectúan pruebas de agregación de resultados con distintos *offsets* y a diferentes tasas de muestreo. Se realizará una descripción del entorno y casos simulados y se mostrarán también los resultados de los informes agregados obtenidos, comparando de los mismos con los de la traza original. Las conclusiones a extraer de todos estos resultados se mostrarán en la sección siguiente.

4.2 Muestreo Convencional

Como ya se mostró en la sección de Caracterización de Tráfico, la extracción de informes *netflow* con muestreo se realiza con distintos parámetros.

El objeto de esta subsección es mostrar los resultados obtenidos mediante muestreo simple, es decir, no existe agregación alguna de los datos obtenidos. Se evalúa, por lo tanto, el funcionamiento de un único nodo de red que implementa un sistema de monitorización *netflow* con muestreo de paquetes a una tasa determinada. Es por lo tanto justo asumir que el *offset* es, en este caso, un factor irrelevante.

4.2.1 Medida de Bondad de Similitud: Chi – Cuadrado y Phi

A la hora de valorar cual es el grado de similitud entre los resultados obtenidos para cada métrica y tasa con aquellos de la traza original al 100%, es necesaria la adopción de un mecanismo que permita dar una medida coherente que sirva como referencia.

En las publicaciones científicas consultadas, como [31], [13], [17] o [14], se puede comprobar cómo el estadístico de prueba adoptado para evaluar esta similitud en informes *netflow* está basado en la prueba *Chi cuadrado* de bondad de ajuste, que se obtiene de manera relativa al resultado de esta. Pasamos a mostrar a continuación las características de la distribución *Chi cuadrado*, el test de similitud asociado a este, y después de la realización de pruebas con el mismo, se mostrará la medida de referencia finalmente utilizada por la comunidad científica y por este estudio.

4.2.1.1 Distribución y Estadístico Chi Cuadrado:

La distribución *Chi cuadrada* viene definida de la siguiente manera:

Sean $Z_1, Z_2 \ldots Z_k$ variables aleatorias distribuidas normal e independientemente, con media $\mu = 0$ y varianza $\sigma^2 = 1$. Entonces, la variable aleatoria

$$X = Z_1^2 + Z_2^2 + \ldots + Z_k^2$$

tiene la función densidad de probabilidad

$$f(x) = \frac{1}{2^{k/2}\Gamma(\frac{k}{2})} x^{(k/2)-1} e^{-x/2}, \text{ para } x > 0$$

y se dice que sigue una distribución Chi-cuadrada con k grados de libertad, lo que se denota como X_k^2.

La forma de esta distribución depende, por tanto, de el valor de k. Para valores pequeños de ésta, la función presenta un claro sesgo hacia la derecha, que va desapareciendo a medida que k aumenta, haciendo la distribución más simétrica hasta que $k \to \infty$, donde la distribución es normal.

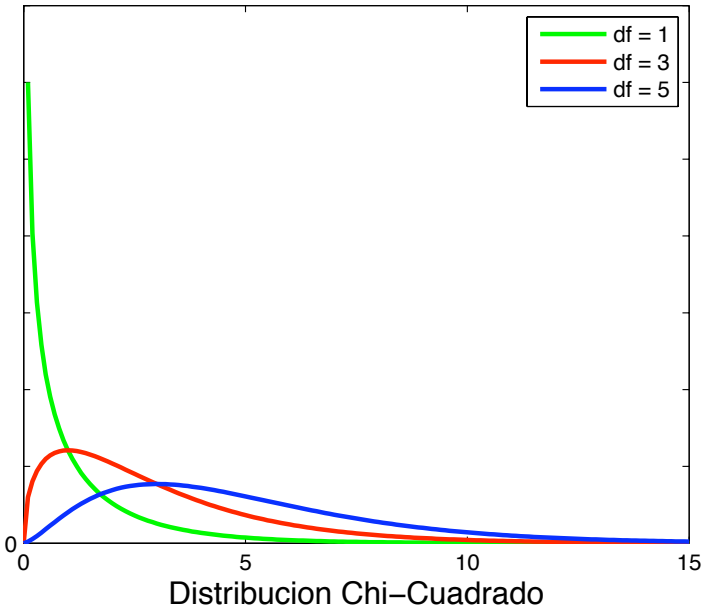


Ilustración 4-1: Distribución Chi Cuadrado según grados de Libertad

Se define un punto crítico en esta distribución como aquel de la variable con *k* grados de libertad tal que la probabilidad de que X sea mayor de este valor es α. Este punto se denota $X_{\alpha,k}^2$:

$$P(X > X_{\alpha,k}^2) = \int_{X_{\alpha,k}^2}^{\infty} f(u)du = \alpha$$

α por lo tanto se trata de un valor de confianza que viene dado por el área bajo la función y el límite inferior $X_{\alpha,k}^2$.

Esta distribución, y el uso de estos valores críticos basados en parámetro de confianza, resulta muy útil a la hora de aplicar el método de bondad de ajuste del mismo nombre.

Esta prueba hace uso de una muestra aleatoria de tamaño *n* proveniente de una distribución desconocida, que se pretende comparar con otra. La muestra se distribuye en sus *n*observaciones en un histograma con k intervalos de clase.

De este modo, si se denota como O_i la frecuencia observada dentro de un intervalo *i*, y E_i la observada en la distribución de referencia dentro del mismo intervalo, el estadístico de la prueba de ajuste es:

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Esta distribución coincide, de manera aproximada, a una distribución Chi-cuadrado *con k-p-1* grados de libertad, siendo *p* el número de parámetros de la distribución propuesta estimada por los estadísticos muestrales.

Se acepta la hipótesis de que la distribución de la población propuesta coincide con la de referencia según el estadístico calculado si

$$X_o^2 > X_{\alpha,k-p-1}^2$$

Un aspecto a tener en cuenta en este test tiene que ver con la magnitud de las frecuencias esperadas, es decir, los intervalos del histograma a aplicar en ambas distribuciones. Si alguna de estas frecuencias es muy pequeña, el estadístico X_o^2 no muestra realmente el alejamiento entre valores esperados y de referencia. Se establecen, por lo tanto, valores mínimos en esta frecuencias para que no quede ningún intervalo despoblado. En [13] y [31], el número mínimo establecido es 5. Esto se consigue fusionando las frecuencias de intervalos adyacentes poco poblados. El proceso entonces pasa por el cálculo de un histograma original para una métrica concreta con un número de cajas o intervalos inicial, en este caso se ha elegido 100. Una vez identificados los intervalos con frecuencias menores que 5, se pasa a unir con el siguiente, obteniendo una distribución con menos intervalos y distribución no uniforme. Se calcula el histograma de frecuencias para ambas poblaciones, de referencia y prueba, con el mismo vector de intervalos. Esto no afecta al resultado final del estadístico, aunque hay que tener en cuenta para la aceptación o rechazo de la hipótesis de pertenencia que los grados de libertad disminuyen. El valor de α adoptado para esta decisión es el típico: 0.05..

4.2.1.2 Distribucionesde Cola Pesada. Estadísticso Phi:

Una de las características del tráfico en Internet, que ya ha sido estudiada, es la aparición de un fenómeno en las distintas medidas practicables a sesiones y flujos de red por el cual estas resultan seguir una distribución de *cola pesada*.

Este fenómeno se traduce en la obtención, para la mayoría de métricas, como por ejemplo, el número de paquetes total, su duración, o el número de bytes intercambiados por flujo, de una distribución que presenta dos características fácilmente apreciables:

- Una es la existencia un claro máximo en valores pequeños, lo que indica que existe una enorme mayoría de flujos cortos.

- La otra característica observable es que, para cualquiera de estas métricas, la distribución no parece llegar a extinguirse nunca, llegando a alcanzar el máximo posible determinado por la longitud de la muestra obtenida.

Existen, por lo tanto, en contraposición con una mayoría de flujos cortos, un número pequeño de flujos que se pueden mantener activos durante mucho tiempo, y que engloban gran cantidad de tráfico. Este efecto se hizo claramente visible en los resultados obtenidos en la sección anterior, de caracterización de la traza original. El uso de un eje de abscisas con distribución logarítmica queda, por lo tanto, claramente justificado.

Este fenómeno ha sido estudiado en distintos casos, como en [38] o en Proyectos Fin de Carrera, como [36]. Estos estudios tratan de caracterizar matemáticamente los resultados obtenidos del análisis de otras trazas de red mediante este tipo de distribución.

El objetivo del presente Proyecto Fin de Carrera pasa por un análisis directo del impacto del muestreo de paquetes en los resultados de informes *netflow* obtenidos con distintas técnicas. Sin embargo, se considera útil la mención de este fenómeno para comprender las limitaciones que el mismo impone a pruebas desimilitud como *Chicuadrado*.

Se muestra ahora, y a modo de ejemplo, cuál es la distribución que se obtiene al realizar un histograma del número de paquetes para la traza completa, muestreada al 100%, y al 50%, para 15 segundos como tiempo máximo entre llegadas:

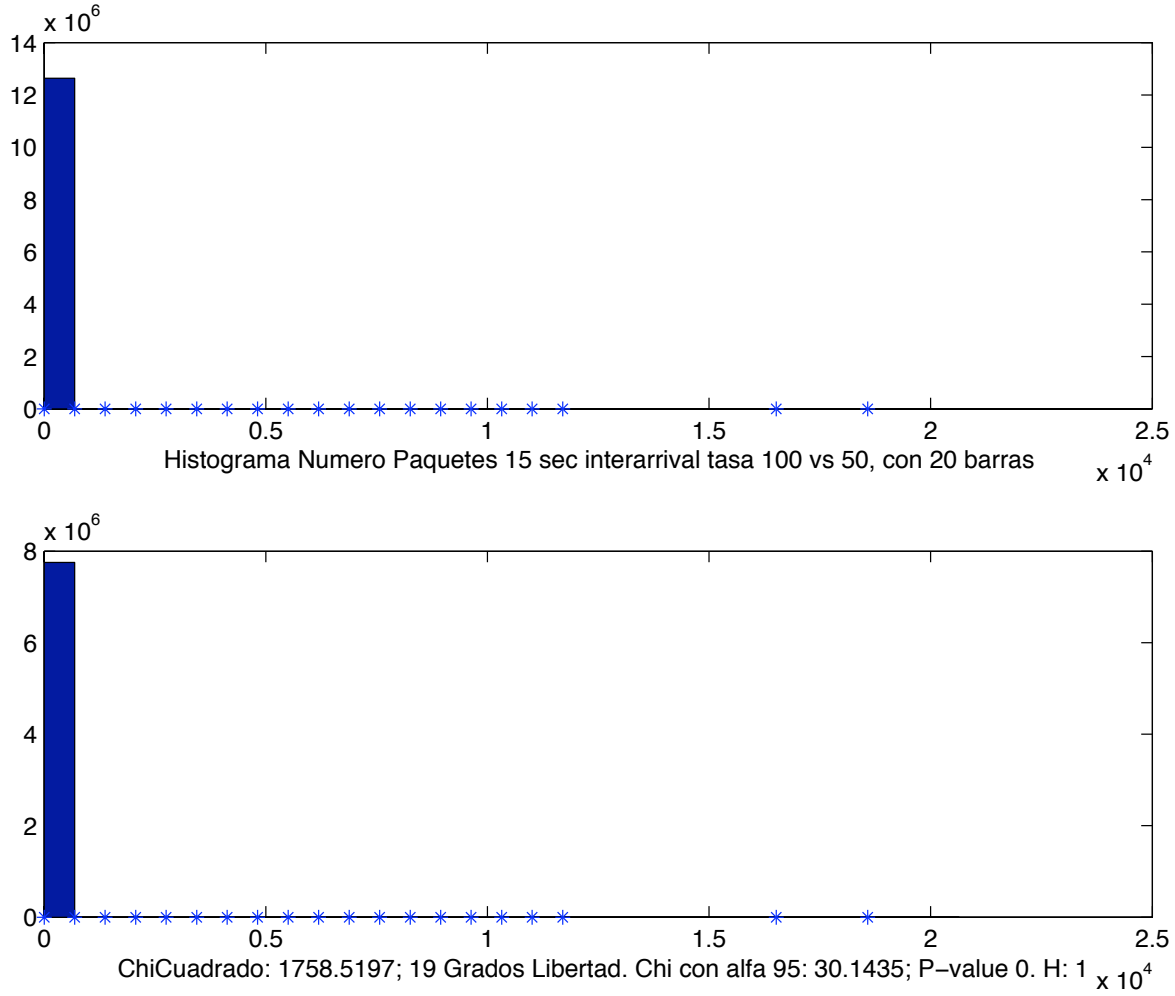


Ilustración 4-2: Histograma ejemplo Numero Paquetes 100% vs 50%

La distribución de los intervalos o cajas del histograma se realiza según lo explicado: se toman los datos de la distribución original y se realiza un histograma preliminar con 100 cajas. Se buscan intervalos que no han sido suficientemente poblados, que se fusionan con el adyacente, y se repite el proceso hasta obtener al menos 5 medidas por bloque. Con ese vector de intervalos se realiza el histograma de los datos de prueba, en este caso al 50%, y se calcula el valor del estadístico Chi-cuadrado.

Una vez hecho esto, se toma el número de intervalos usado para calcular, junto con un valor de confianza de 0.95, el punto crítico de la distribución Chi-cuadrado que se usa como umbral para determinar si ambas distribuciones son distinguibles (H = 1) o no (H = 0).

De la ilustración mostrada se puede observar cómo los resultados obtenidos, que se muestran en las siguientes subsecciones, determinan siempre que ambas distribuciones son, en efecto, absolutamente distinguibles.

Para evitar el efecto de un único intervalo inicial demasiado grande y otros casi vacíos, se hicieron pruebas con distintos parámetros, de número inicial de cajas del histograma y población mínima por intervalo, con resultados muy parecidos. El conocimiento de la existencia del fenómeno de cola pesada indicaba ya, por otro lado, que los resultados obtenidos con esta prueba seguirían esta tendencia.

Los artículos consultados, que hacen estudios parecidos a los de este proyecto, hacen referencia al uso de una medida relativa, basada en *Chi cuadrado,*denominada ***estadístico Phi***, o ϕ.

Este valor, *Phi*, se utiliza como referencia del valor obtenido de la prueba de bondad de ajuste *Chicuadrado* en referencia a la población total de cada una de las pruebas. Proporciona, por lo tanto, un valor mucho más acotado que permite evaluar de manera cuantitativa la pérdida relativa de precisión en las medidas obtenidas con distintas tasa de muestreo.

Su cálculo es sencillo, y deriva directamente del valor obtenido con la prueba de ajuste *Chi cuadrado*. Si este se denotaba por $X_{\alpha,k\text{-}p\text{-}1}{}^2$, y la población total de ambos experimentos sumados, $\sum_i^k E_i + O_i$, es N, el valor de ϕ es:

$$\phi = \sqrt{\frac{X^2_{\alpha,k-p-1}}{N}}$$

A continuación se muestran los resultados obtenidos de la extracción de métricas para cada uno de los experimentos con distintas tasas. Se muestran las gráficas de distribución de probabilidad acumulada siempre junto con la gráfica de referencia al 100%.

Cada métrica se calcula de nuevo para ambos tiempos máximos entre llegadas, y se muestra una gráfica con los resultados del test *Chi cuadrado* y el valor relativo, más significativo, *Phi*.

Por último, y para poder evaluar el factor de degradación de los resultados obtenidos, se muestran gráficas comparando los valores de *Phi* obtenidos para cada tasa de muestreo.

De la evidencia mostrada por las primeras pruebas, que confirman la clara tendencia hacia distribuciones de cola pesada, se puede intuir que los resultados de los tests de similitud estadística obtendrán resultados poco concluyentes cuantitativamente. Esto, por otro lado, no quiere decir que su aplicación no vaya a proporcionar información, ya que serán medidas útiles para hacer una valoración cualitativa entre métodos de muestreo que, unido al análisis más pormenorizado del resto de muestras, puede llevar a conclusiones interesantes que permitan hacer una valoración de las pruebas efectuadas, y la posterior propuesta de un esquema de muestreo conveniente que pueda aportar una mejora a los sistemas existentes.

4.2.1.3 Estimación de resultados a distintas tasas.

A la hora de realizar una comparación de los resultados originales con los obtenidos según tasas, existen métricas que podrían, en cierta medida, permitir una transformación que ayude a contrarrestar el efecto de la obtención de menos muestras (esto es mayor tasa de muestreo).

En la mayoría de los casos de este estudio, esta transformacioón se realiza haciendo una multiplicación de los resultados obtenidos por la inversa de la tasa de muestreo utilizada. Para métricas de volumen, como el número de paquetes o Bytes en un flujo, esta medida es útil.

Sin embargo, esta transformación tiene una limitación, ya que aquellos flujos más cortos se perderán con mayor probabilidad que los más largos en el proceso de muestreo, y por lo tanto no se pueden recuperar de manera convencional. [31] y [32] se hacen eco de este fenómeno y proponen métodos alternativos que explotan algunas características concretas, como la aparición de banderas de inicio (SYN) y finalización (FIN) en las cabeceras de paquetes TCP.

En este caso, por lo tanto, se aplica una transformación directa para las métricas de número de paquetes, tamaño y throughput, pero no para la duración del flujo, ya que esta

multiplicación no aporta en realidad un acercamiento a la medida real. La comparación de la evolución del valor Phi para esta métrica se hace entonces sin transformación alguna.

4.2.2 Duración

En la siguientes gráficas se muestra siempre en azul el resultado obtenido para un muestreo al 100%, y en rojo el resultado para la tasa en cuestión.

La tabla posterior muestra los resultados de la aplicación del test Chi cuadrado y la obtención del estadístico de bondad de ajuste relativo Phi una vez aplicada la transformación correspondiente.

- **15 segundos interarrival:**

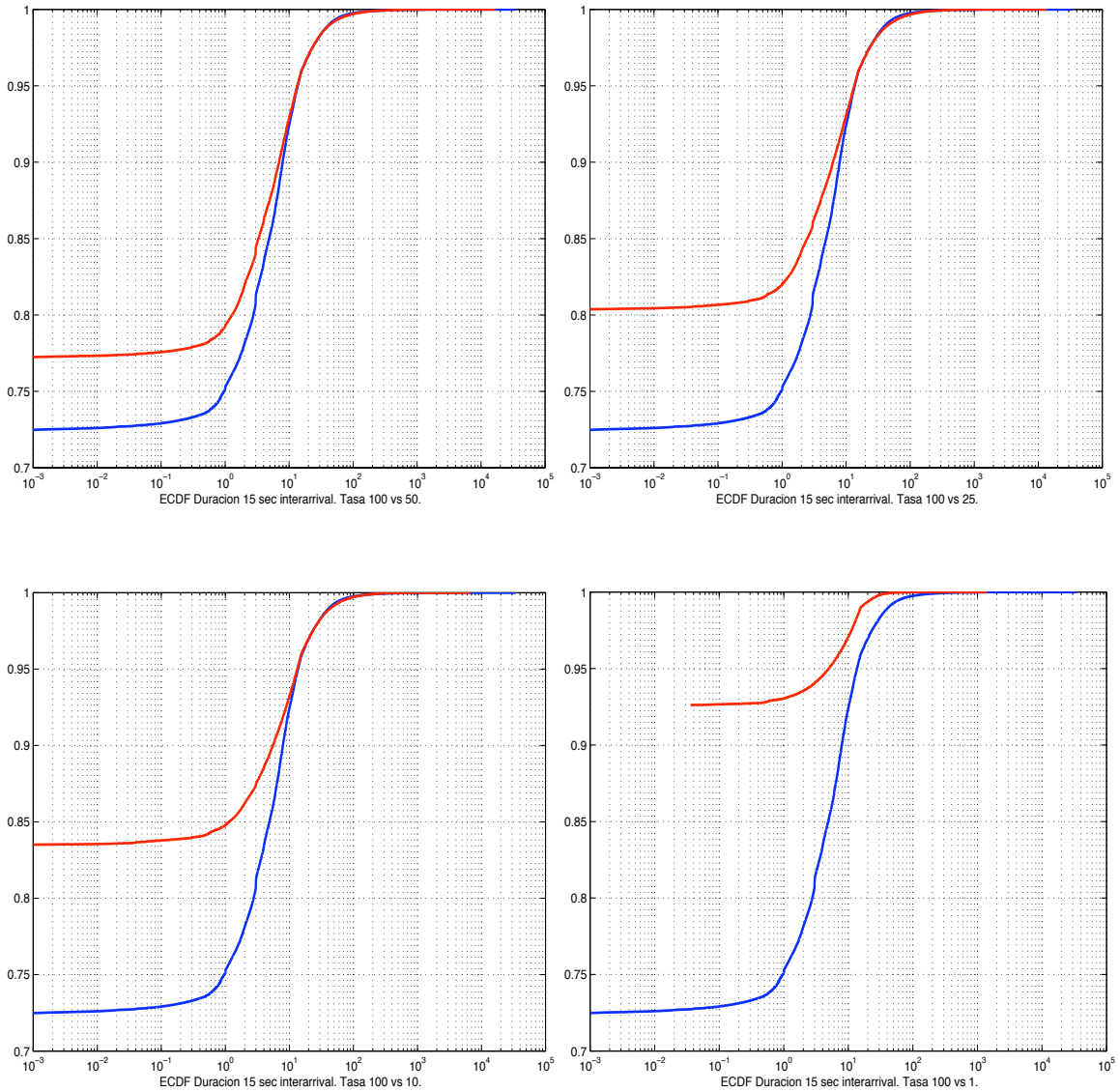


Ilustración 4-3: CDF Duración Muestreado 15s Interarrival

15 interarrival. Alfa = 0.95						
Tasa	Grad.Libertad	Chi Referencia	Chi Test	P value	Hipótesis	Phi
50%	22	33.9244	483.6005	0	1	0.0044
25%	22	33.9244	1229.8417	0	1	0.0070
10%	22	33.9244	1323.3534	0	1	0.0072
1%	22	33.9244	14270.0371	0	1	0.0238

Tabla : Tests Similitud Duracion Muestreado. 15s Interarrival

- **120 segundos interarrival:**

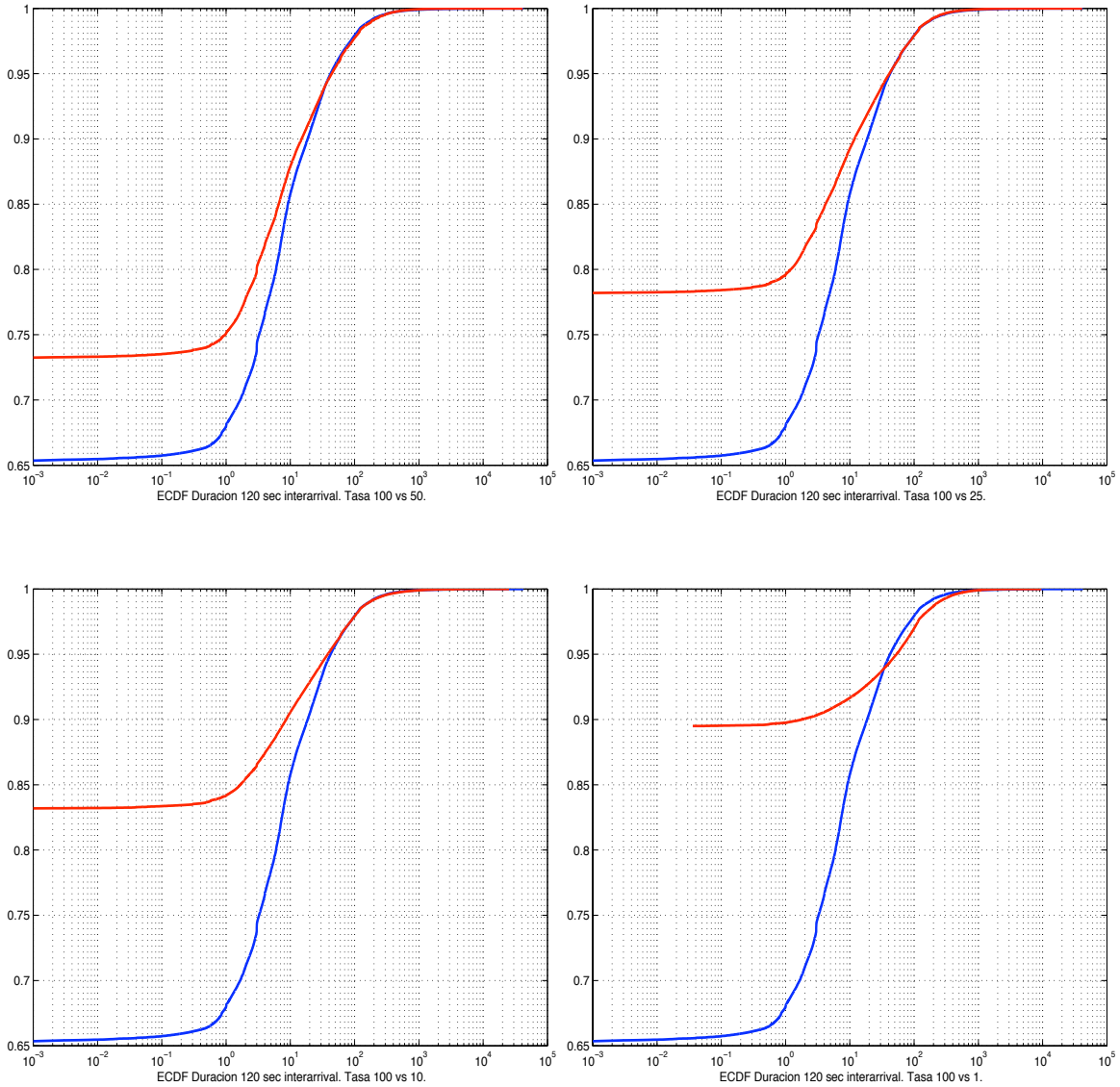


Ilustración 4-4: CDF Duración Muestreado 120s Interarrival

120 interarrival. Alfa = 0.95						
Tasa	Grad.Libertad	Chi Ref.	Chi Test	P value	Hipótesis	Phi
50%	47	64.0011	Inf	0	1	Inf
25%	47	64.0011	Inf	0	1	Inf
10%	47	64.0011	NaN	0	1	NaN
1%	47	64.0011	NaN	0	1	NaN

Tabla 4-1: Tests Similitud Duración Muestreado. 120s Interarrival

Cabe mencionar cómo, para el caso de 120 segundos, las medidas de similitud no son posibles debido a una saturación de las variables obtenidas.

En cualquier caso, se puede observar también la pronunciada diferencia entre funciones de distribución acumuladas que se obtiene al disminuir la tasa de muestro, hasta el punto de deformar el resultado de manera que el número de flujos con valor mínimo, es decir, 0, pasa de un 73% del total para un muestreo de cada 2, a casi un 90% para uno de cada 100 en el caso de 120 segundos entre llegadas, mientras que para 15 segundos, esta diferencia va de 77% para 1 de cada 100 al 93% para 1 de cada 2.

El valor máximo también varía también en ambos casos, aunque no lo hace del mismo modo para cada tiempo entre llegadas. En el caso de 15 segundos la disminución de la duración máxima se hace más palpable al disminuir la tasa de muestreo, mientras que para 120 segundos, esta disminución no es prácticamente distinguible hasta llegar a tasas bajas. Se muestrauna tabla comparativa para ambos:

Valores máximos de Número de paquetes por flujo. 15 y 120 segundos interarrival.					
	100%	50%	25%	10%	1%
15 s.	34595	16801	13479	6805	1402
120 s.	42208	41221	41208	25255	9576

Tabla 4-2: Valores Máximos Duración Muestreado

En definitiva, estos resultados permiten concluir que el muestreo aún a tasas tan bajas como 50% implica que la distribución de la duración de los flujos de red es **estadísticamente distinguible** con la del 100%. Se puede adelantar que este resultado se ha dado en todas las medidas estudiadas en este trabajo.

De este modo toma relevancia el estadístico *Phi* que proporcina una medida cuantitativa de la diferencia entre los distintos experimentos. En concreto la figura 4-5, pone de manifiesto

la tendencia que tiene el estadístico *Phi* para cada tasa. La misma parece indicar que el valor ajuste de un 50% hasta 10% de muestreo no parece variar muy significativamente, mientras que este se dispara al llegar al 1% de tasa.

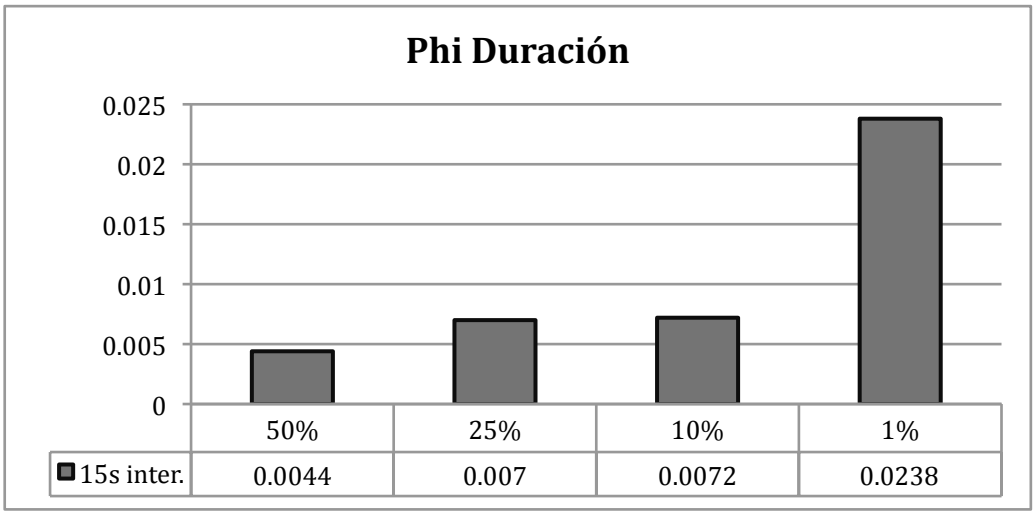
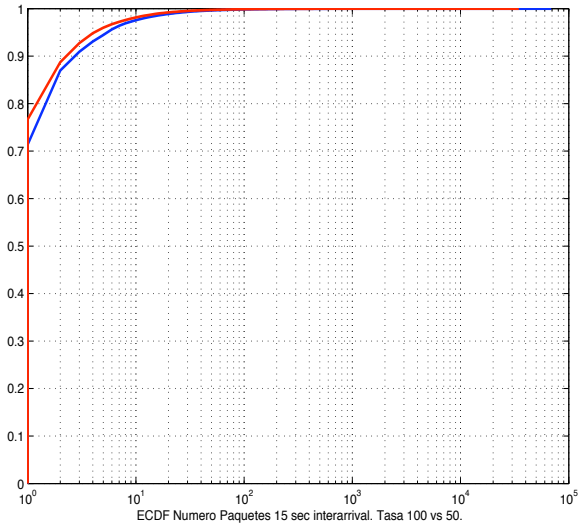
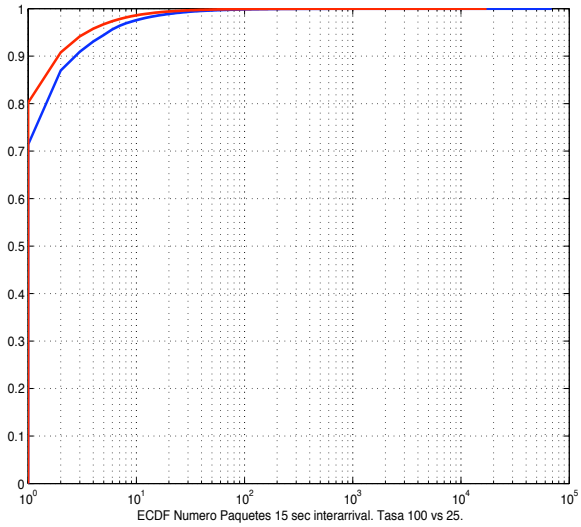


Ilustración 4-5: Duración de Flujos. Evolución Estadístico Phi

4.2.3 Número de Paquetes

- **15 segundos interarrival:**





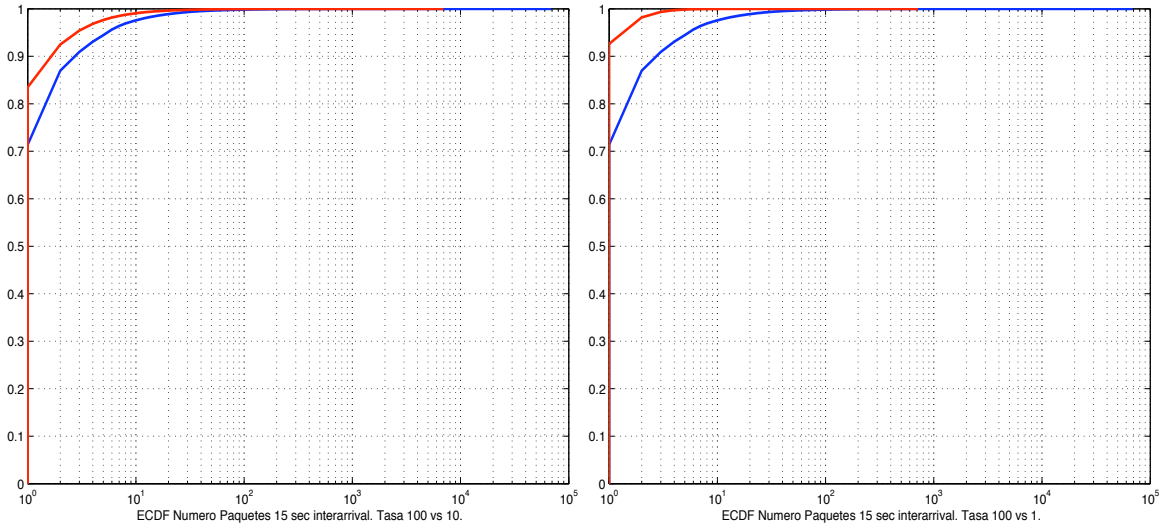
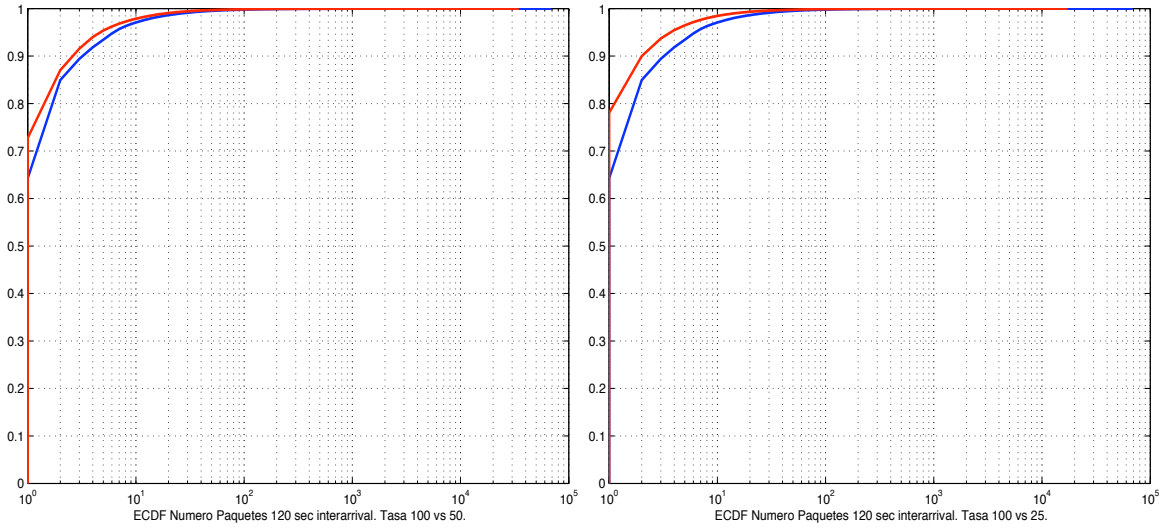


Ilustración 4-6: CDF Número de Paquetes Muestreado 15s Interarrival

15 interarrival. Alfa = 0.95						
Tasa	Grad.Libertad	Chi Ref.	Chi Test	P value	Hipótesis	Phi
50%	19	30.1435	1758.5197	0	1	0.0083
25%	19	30.1435	10490.2083	0	1	0.0204
10%	19	30.1435	67829.1817	0	1	0.0518
1%	19	30.1435	1516189.88	0	1	0.2448

Tabla 4-3: Tests Similitud Número Paquetes Muestreado. 15s Interarrival

- **120 segundos interarrival:**



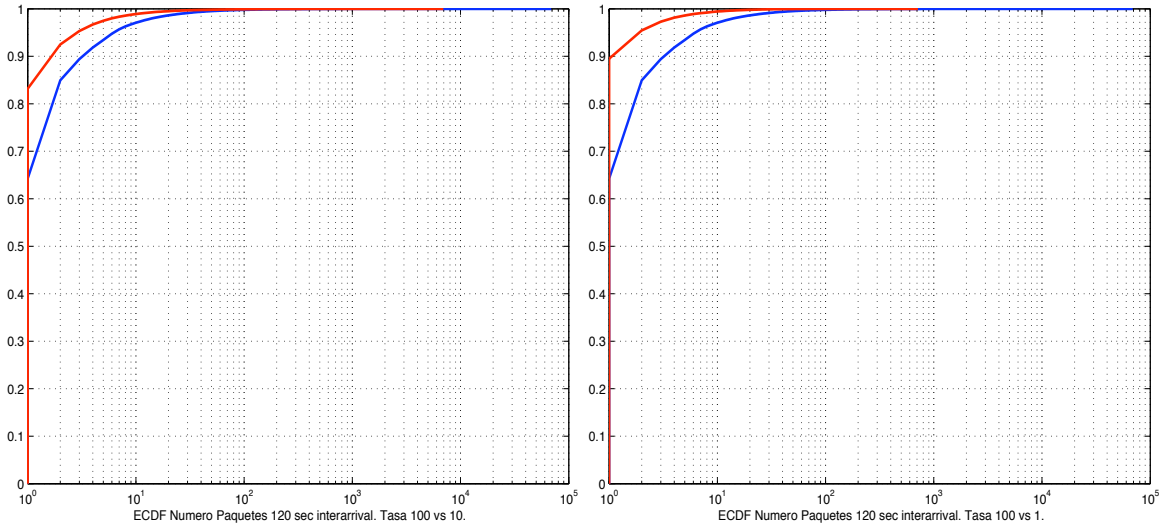


Ilustración 4-7: CDF Número de Paquetes Muestreado 120s Interarrival

120 interarrival. Alfa = 0.95						
Tasa	Grad.Libertad	Chi Ref.	Chi Test	P value	Hipótesis	Phi
50%	23	35.1725	1647.3083	0	1	0.0088
25%	23	35.1725	13484.1508	0	1	0.0251
10%	23	35.1725	105310.9791	0	1	0.0702
1%	23	35.1725	15570175.25	0	1	0.8541

Tabla 4-4: Tests Similitud Número Paquetes Muestreado. 120s Interarrival

Se hace patente cómo, para ambos casos de tiempo entre llegadas, la disminución en la tasa de muestreo provoca un aumento gradual del número de flujos con un único paquete.

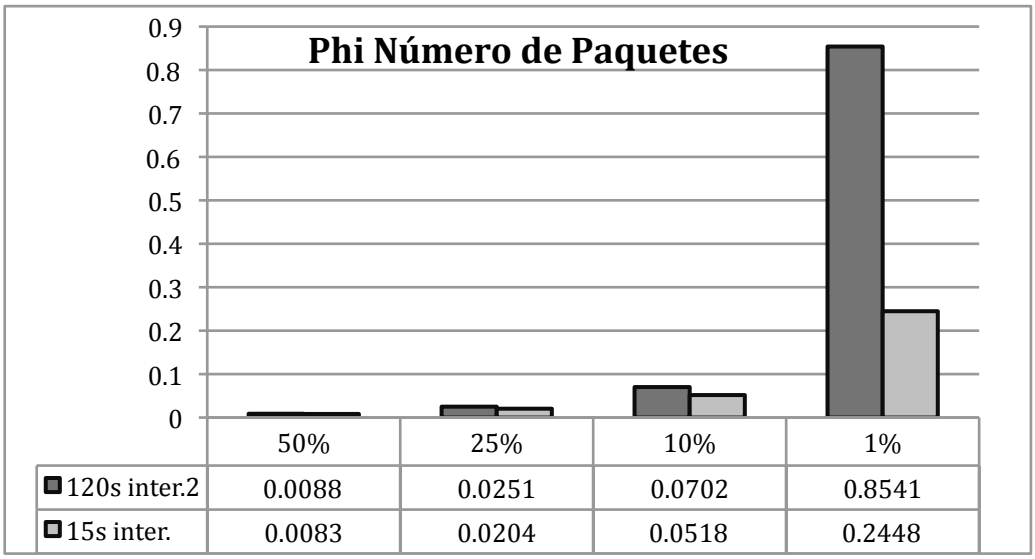
A diferencia que con la métrica anterior, de duración, y por inspección visual, estas diferencias en el porcentaje de flujos con el valor mínimo no llegan a ser tan grandes como antes. La diferenciación estadística con los resultados de la traza original, sin embargo, sigue siendo una constante clara, como indican los resultados del test *Chi cuadrado*.

Los valores máximos alcanzados por cada una de las ejecuciones varían en relación directa con la tasa de muestreo utilizada, lo que es de esperar. Sin embargo, cabe destacar cómo estos no varían en absoluto para distintos tiempos entre llegadas.

Valores máximos de Número de paquetes por flujo. 15 y 120 segundos interarrival.				
100%	50%	25%	10%	1%
68812	34533	17233	6936	714

Tabla 4-5: Valores Máximos Número Paquetes Muestreado

De los resultados obtenidos con las pruebas bondad de ajuste, se extrae que este valor sigue una tendencia lineal con la tasa, pero que para 15 segundos, estos valores son más limitados y por lo tanto más ajustados a la distribución real.



	50%	25%	10%	1%
120s inter.2	0.0088	0.0251	0.0702	0.8541
15s inter.	0.0083	0.0204	0.0518	0.2448

Ilustración 4-8: Número de Paquetes. Evolución Estadístico Phi

En el caso de querer obtener el máximo de información de todos los flujos, es clara la tendencia que siguen los resultados, pero se considera interesante apuntar que aunque tanto una inspección visual de los resultados para ambos tiempos, como los resultados obtenidos para el estadístico de bondad de ajuste Phi, evidencian que la distribuciones no son iguales, el hecho de que los flujos más grandes serán detectados de manera casi idéntica, puede indicar que esta decisión de diseño no es tan crítica como cabría esperar si el objetivo de la monitorización se centrara en captar únicamente flujos grandes. Estos resultados estarían en la línea de [20], y en éste sentido indican que la adopción de tiempos entre llegadas distintos para una solución de éste tipo, que únicamente se centre en grande flujos, es relativamente poco relevante para estas métricas.

4.2.4 Tamaño

El tamaño, como ya se explicó en el anterior capítulo, hace referencia a la agregación del número de bytes de la carga útil de los paquetes pertenecientes a cada flujo. No se tienen en cuenta los paquetes con carga nula, por lo que el valor mínimo en las gráficas de distribución será 1.

- **15 segundos interarrival:**

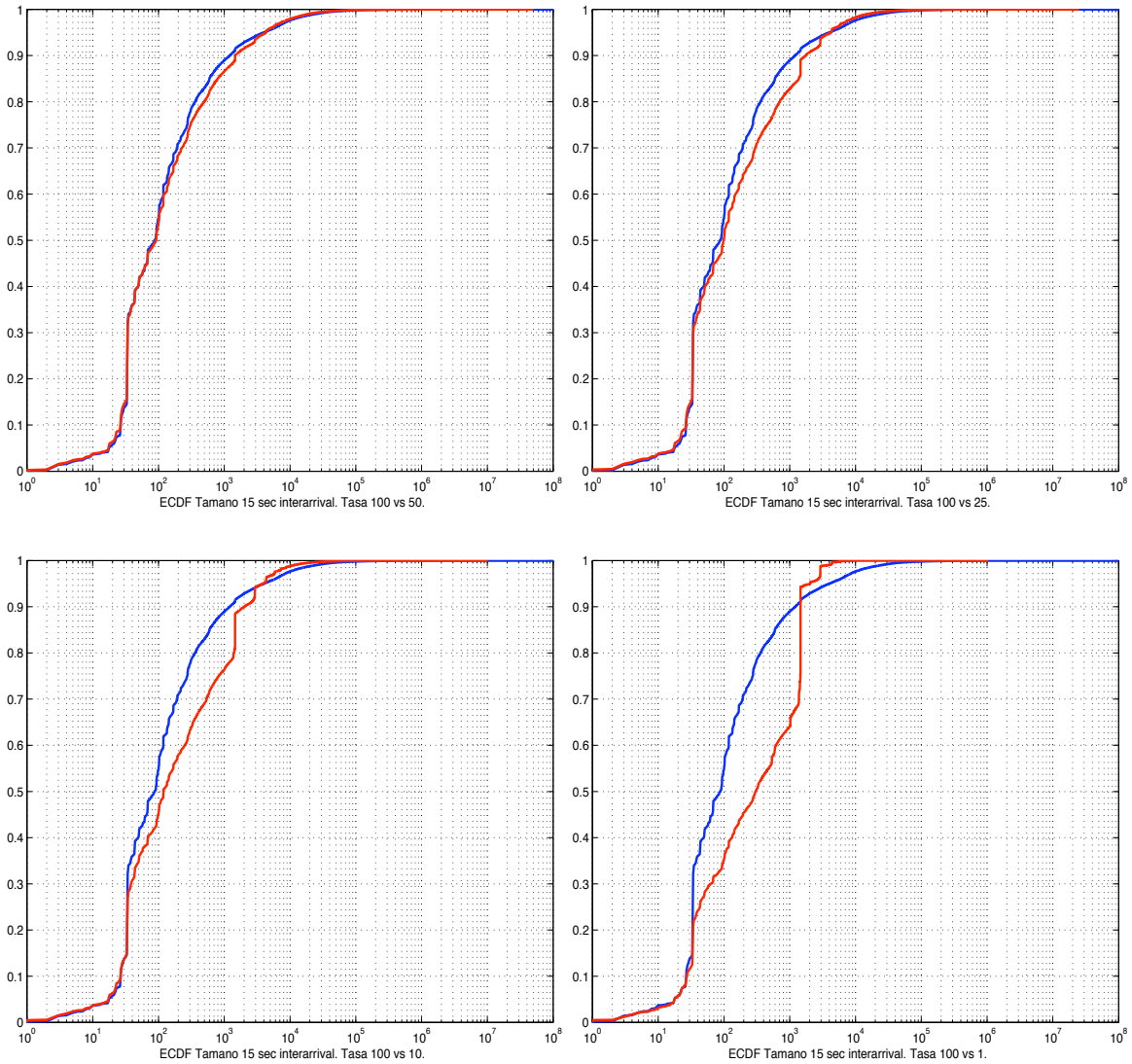


Ilustración 4-9: CDF Número de Bytes Muestreado 15s Interarrival

15 interarrival. Alfa = 0.95						
Tasa	Grad.Libertad	Chi Ref.	Chi Test	P value	Hipótesis	Phi
50%	15	24.9958	1147.2939	0	1	0.0067
25%	15	24.9958	6897.7809	0	1	0.0165
10%	15	24.9958	45316.7878	0	1	0.0423
1%	15	24.9958	443230.3603	0	1	0.1324

Tabla 4-6: Tests Similitud Número Bytes Muestreado. 15s Interarrival

- **120 segundos interarrival:**

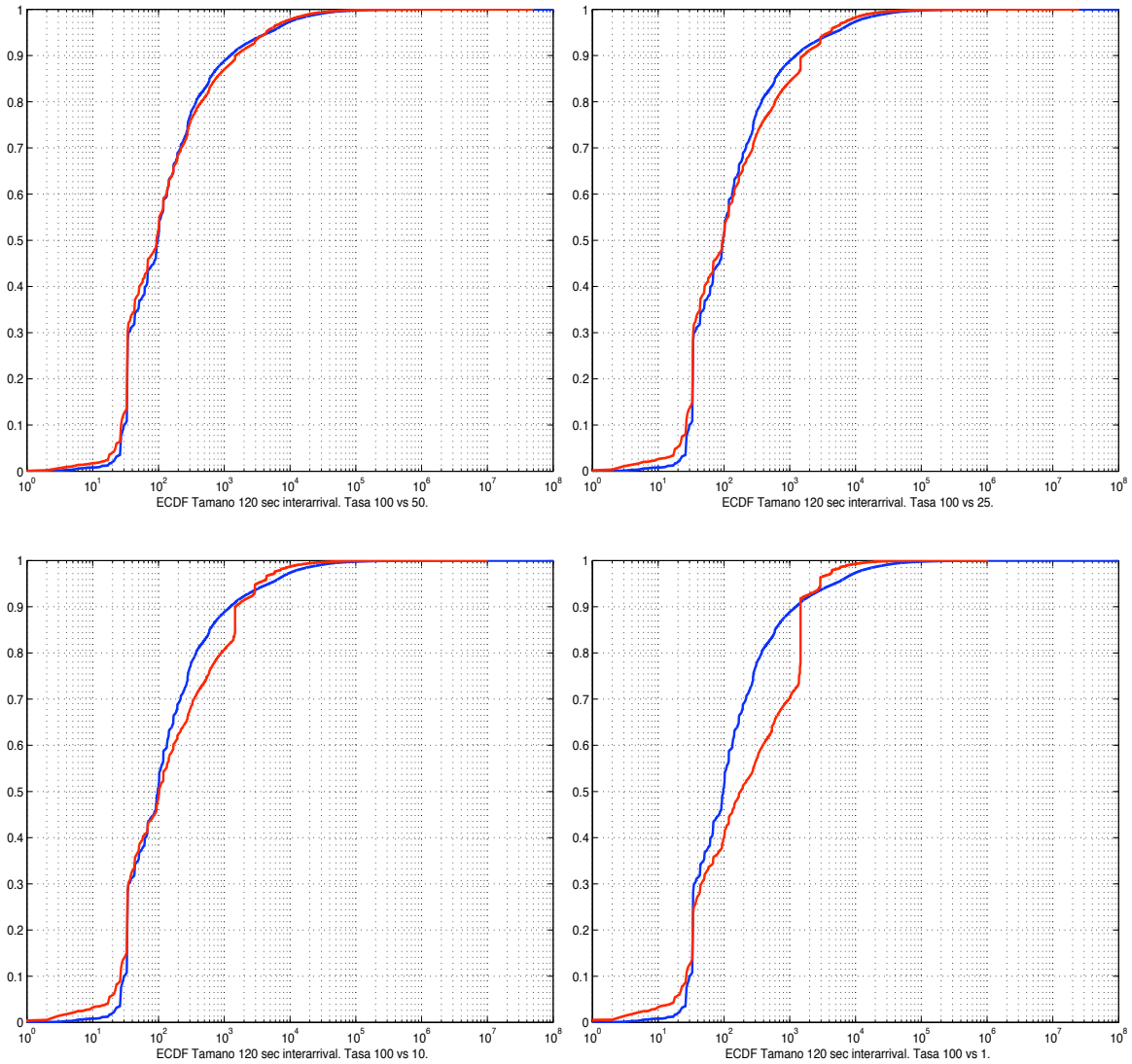


Ilustración 4-10: CDF Número de Bytes Muestreado 120s Interarrival

120 interarrival. Alfa = 0.95						
Tasa	Grad.Libertad	Chi Ref.	Chi Test	P value	Hipótesis	Phi
50%	17	27.5871	1317.5839	0	1	0.0079
25%	17	27.5871	10719.6818	0	1	0.0224
10%	17	27.5871	83331.3147	0	1	0.0625
1%	17	27.5871	10660813.53	0	1	0.7068

Tabla 4-7: Tests Similitud Número Bytes Muestreado. 120s Interarrival

Esta métrica tiene el mismo comportamiento que la del número de paquetes en cuanto a sus valores máximos, por lo que estos son los mismos tanto para 15 como para 120 segundo de interarrival, y tienden a disminuir de manera lineal con la tasa de muestreo.

Valores máximos de Bytes por flujo. 15 y 120 segundos interarrival.				
100%	50%	25%	10%	1%
100462777	50416965	25158965	10126560	1042440

Tabla 4-8: Valores Máximos Bytes por Flujo Muestreado

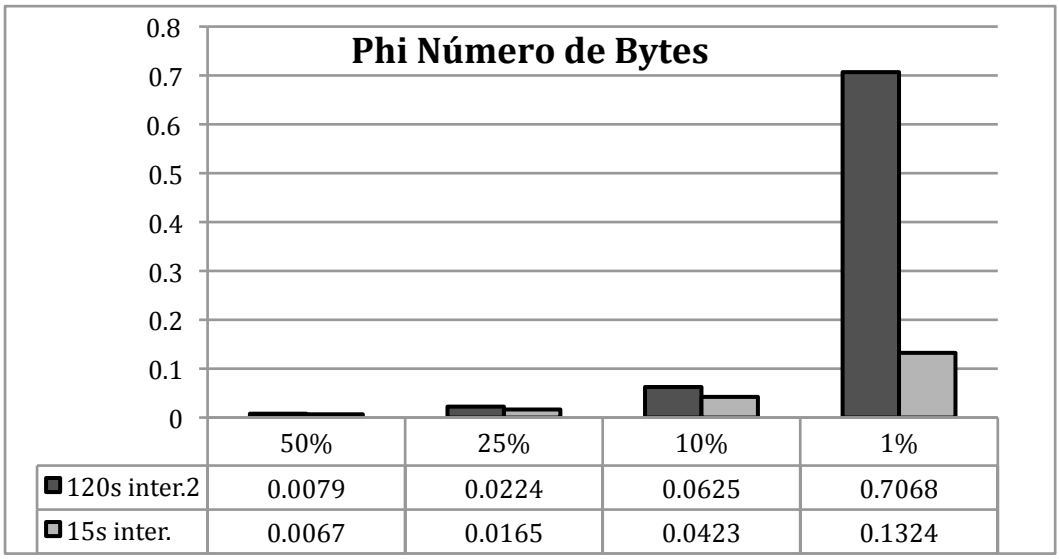
Otra característica apreciable es al aparición de una nueva moda en las distribuciones, que se acentúa con la disminución de nuevo de la tasa de muestreo. Si bien la moda ya identificada de la traza original, que según análisis correspondería con peticiones DNS y que está entorno a los 34 bytes, sigue apareciendo, esta disminuye levemente su impacto, mientras que otra, centrada en los 1460 bytes, crea un marcado escalón, muy apreciable especialmente en los casos de menor tasa.

Para averiguar el origen de estos flujos se vuelve a hacer un análisis manual de las características de los flujos correspondientes a este tamaño para cada tasa de muestreo. De este análisis se extrae que la mayoría de estos flujos contienen un único paquete y corresponden a conexiones TCP al puerto 80, a direcciones IP correspondientes a dominios muy dispares.

Se pasa entonces a hacer una valoración del protocolo TCP. Este cuenta con un mecanismo para establecer el tamaño máximo de segmento (MSS) entre equipos. Si esta coordinación no se realiza, este tamaño es de 536 bytes, lo que resulta en un pobre rendimiento. En el caso de que ambos hosts estén dentro de la misma red, esta indicación de tamaño sí se realiza, y viene condicionado por el valor máximo de transmisión (MTU) a nivel Ethernet. Si se tiene en cuenta que éste es de 1500 bytes, y tras un substracción de 20 bytes para la cabecera IP y otros 20 para TCP, queda un MSS de 1460 bytes.

Se considera, por tanto, que el origen de esta moda en el tamaño de flujos se debe al aumento en la aparición de flujos con un único paquete, cuya tendencia crece al disminuir la tasa de muestreo, como se puede observar en el apartado anterior, y su tamaño está determinado por el máximo establecido a nivel de enlace, lo que da lugar a paquetes de 1460 bytes de payload.

Por último se muestra una tabla comparativa con los resultado del estadístico *Phi*, que sigue la misma tendencia que para las métricas anteriores, siendo estos valores más pequeños para 15 segundos entre llegadas.



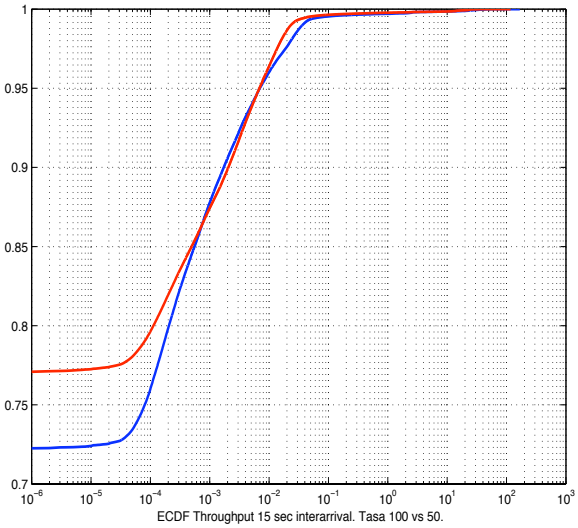
	50%	25%	10%	1%
120s inter.2	0.0079	0.0224	0.0625	0.7068
15s inter.	0.0067	0.0165	0.0423	0.1324

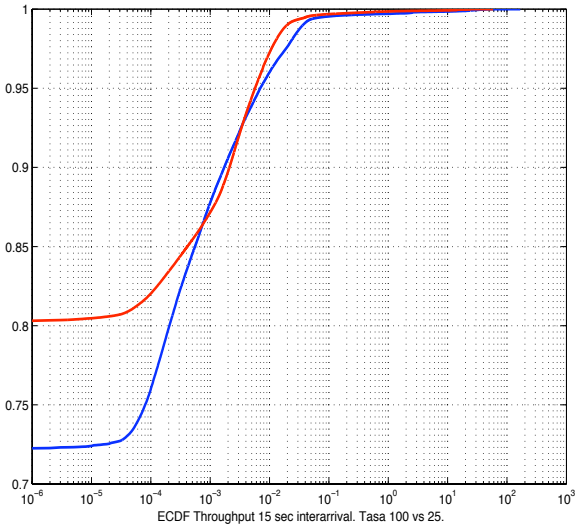
Ilustración 4-11: Número de Bytes. Evolución Estadístico Phi

4.2.5 Throughput

Se muestran los resultados obtenidos para la medida del Throughput, o caudal empleado por los flujos detectados.

- **15 segundos interarrival:**





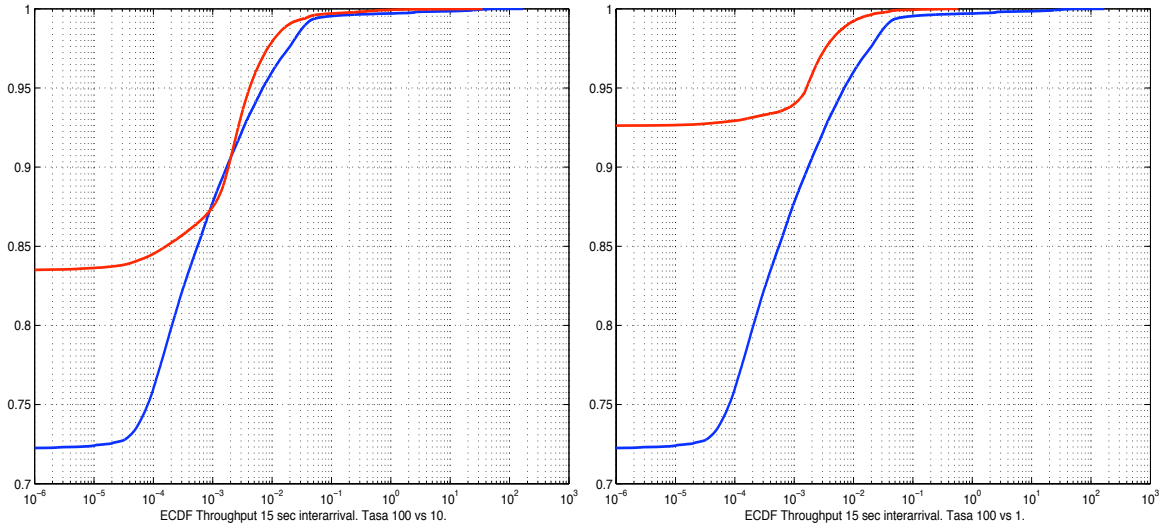
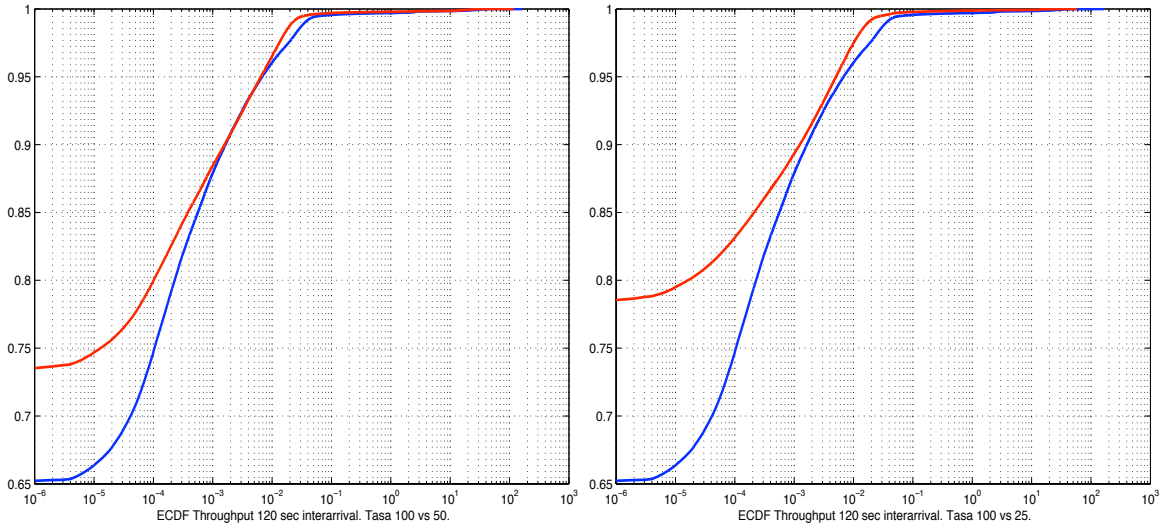


Ilustración 4-12: CDF Throughput Muestreado 15s Interarrival

15 interarrival. Alfa = 0.95						
Tasa	Grad.Libertad	Chi Ref.	Chi Test	P value	Hipótesis	Phi
50%	64	83.6753	44999.4476	0	1	0.0422
25%	64	83.6753	630504.351	0	1	0.1579
10%	64	83.6753	129656.0859	0	1	0.0716
1%	64	83.6753	680476.8314	0	1	0.164

Tabla 4-9: Test similitud Throughput Muestreado. 15s Interarrival

- **120 segundos interarrival:**



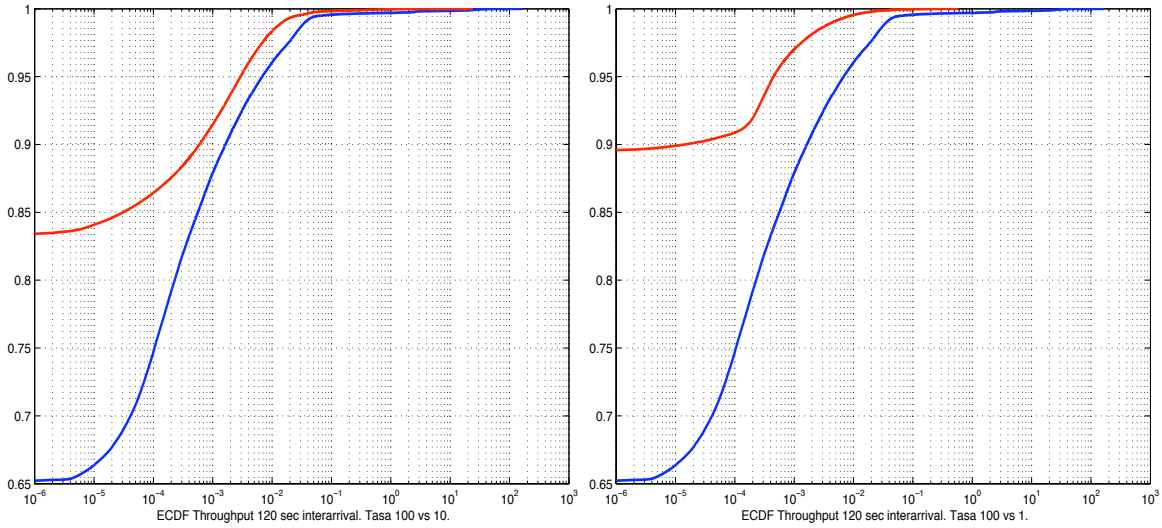


Ilustración 4-13: CDF Throughput Muestreado 120s Interarrival

120 interarrival. Alfa = 0.95						
Tasa	Grad.Libertad	Chi Ref.	Chi Test	P value	Hipótesis	Phi
50%	57	75.6237	35093.6947	0	1	0.0406
25%	57	75.6237	335122.3274	0	1	0.1253
10%	57	75.6237	40406.3864	0	1	0.0435
1%	57	75.6237	165383.4056	0	1	0.088

Tabla 4-10: Tests Similitud Throughput. 120s Interarrival

Se hace patente, al igual que para el caso de duración, cómo existe una tendencia uniforme en ciertos rangos de las distribuciones (tras las transformación logarítmica), aunque el tamaño de este intervalo disminuye claramente al disminuir la tasa. De nuevo, la diferencia en el porcentaje de flujos que tienen valor mínimo se hace evidente con respecto a la traza original. Dado que existen muchos flujos con un único paquete, y por lo tanto con duración 0, aparecen en la misma medida caudales nulos, lo que pone en tela de juicio la bondad de esta métrica, y la posibilidad de extraer conclusiones directas sobre las características concretas del tráfico a través de las mismas.

En este caso los valores máximos entre ambos tiempos entre llegadas llegan a variar muy ligeramente en algunos casos, pero en su mayoría se mantienen iguales, como pone de manifiesto la siguiente tabla:

Valores máximos de Caudal por flujo. 15 y 120 segundos interarrival.				
100%	50%	25%	10%	1%
166,7401	116,8085	58,4042	35,0425	0,5811
166,4285	116,8085	58,4042	24,0897	0,5811

Tabla 4-11: Valores Máximos Throughput Muestreado

Los resultados obtenidos con los tests de bondad de ajuste tienen una variabilidad poco clara, por lo que no se pueden extraer conclusiones claras sobre tu tendencia. Cabe mencionar, sin embargo, que si bien la tendencia parece ser la misma para ambos tiempos, es la única métrica para la que los valores extraídos con 15 segundos entre llegadas son sensiblemente peores.

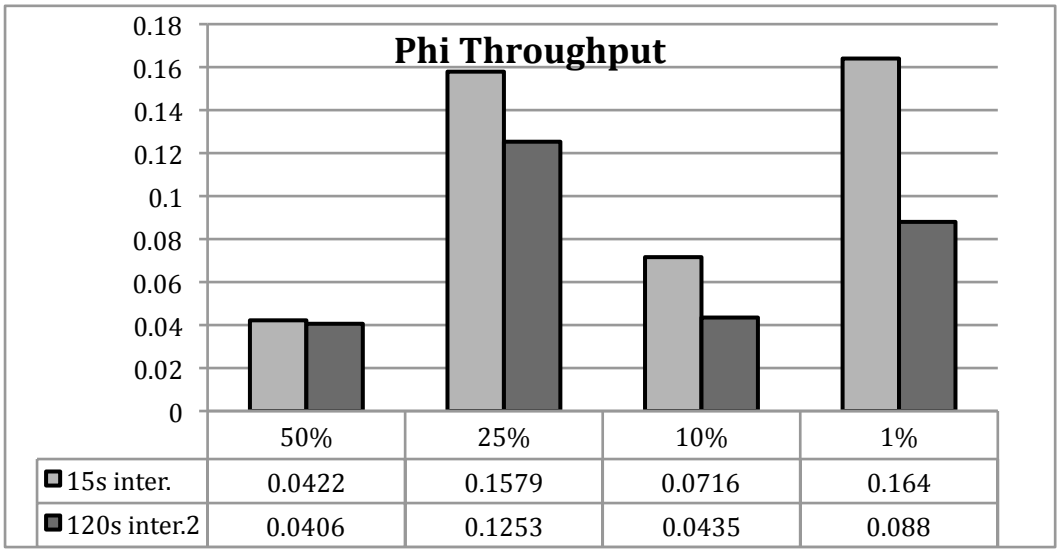


Ilustración 4-14: Throughput. Evolución Estadístico Phi

4.2.6 Porcentaje por Aplicaciones

En este apartado se mostrarán las listas con la distribución de la cantidad de tráfico generada por las distintas aplicaciones detectadas por el generador de flujos implementado. A continuación de estas se mostrarán las gráficas de distribución de aplicaciones agregadas en grupos.

- **Tasa 50%:**

15s interarrival			120 s interarrival		
Application	**Bytes**	**Percentage**	**Application**	**Bytes**	**Percentage**
UnKnown	6137358570	50.9	http	5303513394	43.98
http	4710894072	39.07	UnKnown	5290029305	43.87
Several	477512887	3.96	Several	553389561	4.59
edonkey	280916300	2.33	edonkey	464645265	3.85

ssl	236332692	1.96	ssl	274853336	2.28
pop3	144103512	1.2	pop3	144103512	1.2
msnmessenger	88026982	0.73	bittorrent	100581717	0.83
smtp	85730832	0.71	msnmessenger	87615916	0.73
dns	77223969	0.64	smtp	85733519	0.71
socks	69321934	0.57	dns	77223871	0.64
bittorrent	57258826	0.47	socks	74226634	0.62
rtsp	45426650	0.38	rtsp	45429030	0.38
nbns	34909239	0.29	nbns	24727385	0.21
skypetoskype	22234009	0.18	skypetoskype	22959755	0.19
rtp	15813723	0.13	rtp	13315438	0.11
fasttrack	13331258	0.11	fasttrack	12326257	0.1
stun	10040334	0.08	freenet	8439333	0.07
xunlei	7582530	0.06	xunlei	7477146	0.06
pplive	5737042	0.05	stun	4838825	0.04
smb	3139753	0.03	smb	3236285	0.03
netbios	2566595	0.02	netbios	2470063	0.02
napster	2319186	0.02	pplive	2136046	0.02
yahoo	1932789	0.02	yahoo	2006337	0.02
qq	1097419	0.01	qq	1137066	0.01
imap	1053096	0.01	imap	1083264	0.01
soulseek	1025169	0.01	soulseek	1035007	0.01
ftp	851985	0.01	napster	895062	0.01
freenet	816655	0.01	ftp	851985	0.01
marca	269310	0	vnc	310375	0
x11	116076	0	marca	269310	0
vnc	115956	0	sip	122306	0
sip	115409	0	armagetron	112483	0
gnutella	113595	0	youTubeClick	97824	0
armagetron	105449	0	megaupload	90859	0
megaupload	90859	0	x11	80833	0
youTubeClick	87516	0	elMundo	77275	0
elMundo	77275	0	abc	73146	0
abc	72950	0	gnutella	67947	0
rlogin	65362	0	rdp	60990	0
rdp	44624	0	rlogin	56157	0
megauploadAcceso	35031	0	megauploadAcceso	35766	0
nntp	32089	0	nntp	32089	0
ntp	29750	0	ntp	29750	0
jabber	24225	0	jabber	25003	0
elPais	18522	0	elPais	18522	0
rapidshareAcceso	12439	0	rapidshareAcceso	12439	0
h323	8458	0	irc	8555	0
irc	5766	0	h323	8458	0
aim	2044	0	aim	7150	0
rapidshare	1852	0	rapidshare	1852	0
lpd	777	0	lpd	777	0
shoutcast	731	0	shoutcast	731	0

ventrilo	228	0	ventrilo	228	0
imesh	200	0	imesh	200	0
gopher	129	0	pcanywhere	2	0
battlefield2142	35	0	**TOTAL**	**12058491780**	**100**
pcanywhere	2	0			
TOTAL	**12058491780**	**100**			

Tabla 4-12: Porcentaje por aplicaciones Muestreado 50%

- **Tasa 25%:**

15s interarrival			**120s interarrival**		
Application	**Bytes**	**Percentage**	**Application**	**Bytes**	**Percentage**
UnKnown	3904630136	64.76	UnKnown	3441823528	57.08
http	1597017191	26.49	http	1937926582	32.14
Several	143905295	2.39	edonkey	211790525	3.51
edonkey	128412561	2.13	Several	183491023	3.04
pop3	71982475	1.19	ssl	82930401	1.38
ssl	65086331	1.08	pop3	71982475	1.19
msnmessenger	44372809	0.74	msnmessenger	44158586	0.73
smtp	42934599	0.71	smtp	42935769	0.71
dns	38506744	0.64	bittorrent	41441861	0.69
bittorrent	28171922	0.47	dns	38506744	0.64
socks	27445076	0.46	socks	32505101	0.54
skypetoskype	16161796	0.27	rtsp	21660447	0.36
rtsp	14699684	0.24	skypetoskype	13237850	0.22
nbns	9228942	0.15	nbns	11725094	0.19
fasttrack	8874546	0.15	fasttrack	7706884	0.13
rtp	8824352	0.15	rtp	7148533	0.12
pplive	5809354	0.1	pplive	6264654	0.1
stun	5530591	0.09	xunlei	3776793	0.06
xunlei	3813410	0.06	stun	2970479	0.05
netbios	1621504	0.03	netbios	1576379	0.03
smb	1295019	0.02	smb	1340144	0.02
napster	1196461	0.02	freenet	1141259	0.02
yahoo	955104	0.02	qq	1047283	0.02
qq	739803	0.01	napster	1014125	0.02
soulseek	641310	0.01	yahoo	994556	0.02
ftp	413198	0.01	soulseek	624804	0.01
x11	238659	0	ftp	413198	0.01
marca	145005	0	marca	145005	0
freenet	136929	0	vnc	118083	0
armagetron	122055	0	armagetron	116279	0
h323	87002	0	h323	87002	0
rlogin	71841	0	x11	60537	0
aim	59957	0	aim	59521	0
sip	54359	0	sip	55599	0
gnutella	52527	0	megaupload	43601	0
megaupload	43601	0	rlogin	39750	0

youTubeClick	40063	0	youTubeClick	36357	0
elMundo	36317	0	elMundo	36317	0
abc	34298	0	abc	35768	0
vnc	31528	0	gnutella	31665	0
nntp	26282	0	nntp	26282	0
imap	24633	0	imap	26188	0
megauploadAcceso	18736	0	rdp	21519	0
ntp	14916	0	megauploadAcceso	18736	0
rdp	14470	0	ntp	14916	0
elPais	8980	0	elPais	8980	0
rapidshareAcceso	8409	0	rapidshareAcceso	8409	0
jabber	5298	0	jabber	5435	0
irc	2479	0	irc	3047	0
rapidshare	771	0	rapidshare	771	0
lpd	402	0	lpd	402	0
shoutcast	292	0	gopher	341	0
gopher	129	0	shoutcast	292	0
imesh	100	0	imesh	100	0
pcanywhere	2	0	pcanywhere	2	0
TOTAL	**6029644958**	**100**	**TOTAL**	**6029644958**	**100**

Tabla 4-13: Porcentaje por Aplicaciones Muestreado 25%

- **Tasa 10%:**

15s interarrival			120s interarrival		
Application	**Bytes**	**Percentage**	**Application**	**Bytes**	**Percentage**
UnKnown	1817938355	75.36	UnKnown	1667316460	69.12
http	388212187	16.09	http	512064837	21.23
edonkey	51550493	2.14	edonkey	65362315	2.71
pop3	28906144	1.2	Several	39325196	1.63
Several	28001852	1.16	pop3	28906144	1.2
msnmessenger	17707132	0.73	ssl	20525749	0.85
smtp	17130281	0.71	msnmessenger	17664258	0.73
ssl	16116104	0.67	smtp	17130608	0.71
dns	15449931	0.64	nbns	16208882	0.67
nbns	13487863	0.56	dns	15450004	0.64
bittorrent	11196244	0.46	bittorrent	12210297	0.51
skypetoskype	8355921	0.35	socks	9386465	0.39
socks	7668872	0.32	rtsp	8638442	0.36
fasttrack	4046276	0.17	skypetoskype	5601020	0.23
rtp	3411875	0.14	rtp	3544611	0.15
rtsp	2912730	0.12	fasttrack	3484888	0.14
stun	1757236	0.07	stun	1843496	0.08
xunlei	1502857	0.06	xunlei	1502574	0.06
pplive	1420322	0.06	pplive	1358651	0.06
netbios	732678	0.03	qq	1097835	0.05
napster	539983	0.02	netbios	722518	0.03
smb	424241	0.02	smb	434401	0.02

qq	404033	0.02	yahoo	378428	0.02
yahoo	377422	0.02	napster	370379	0.02
soulseek	295042	0.01	soulseek	314335	0.01
ftp	190393	0.01	freenet	195466	0.01
x11	164017	0.01	ftp	190393	0.01
armagetron	78870	0	x11	76807	0
marca	49732	0	marca	49732	0
rlogin	44857	0	rlogin	27134	0
sip	23508	0	sip	23769	0
gnutella	21615	0	armagetron	22361	0
abc	18661	0	gnutella	22336	0
youTubeClick	16283	0	youTubeClick	20799	0
megaupload	16122	0	abc	18752	0
freenet	14332	0	megaupload	16122	0
elMundo	12788	0	elMundo	12788	0
nntp	10220	0	nntp	10220	0
megauploadAcceso	9056	0	megauploadAcceso	9056	0
ntp	6181	0	ntp	6181	0
h323	5246	0	rdp	5773	0
elPais	5018	0	h323	5246	0
rdp	4789	0	elPais	5018	0
rapidshareAcceso	2381	0	rapidshareAcceso	2381	0
imap	1766	0	imap	1976	0
irc	1115	0	irc	1265	0
ventrilo	201	0	ventrilo	201	0
jabber	181	0	jabber	181	0
lpd	165	0	lpd	165	0
shoutcast	146	0	shoutcast	146	0
gopher	73	0	gopher	73	0
TOTAL	**2412241938**	**100**	**TOTAL**	**2412241938**	**100**

Tabla 4-14: Porcentaje por Aplicaciones Muestreado 10%

- **Tasa 1%:**

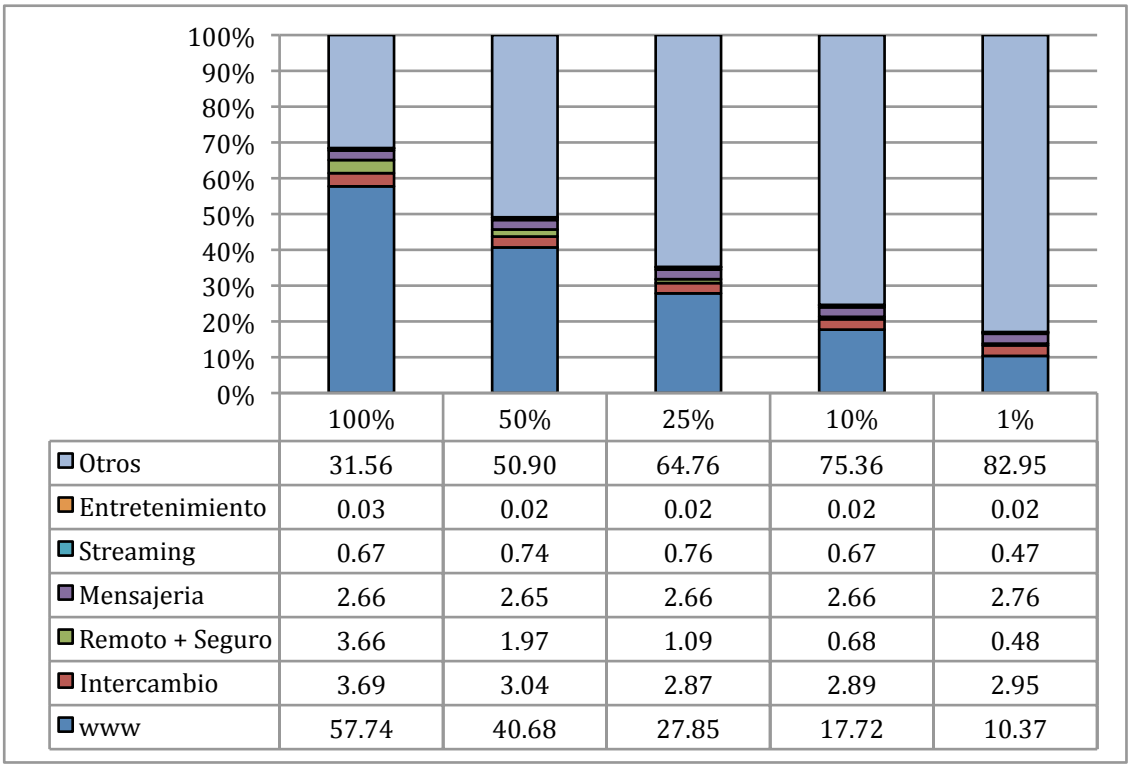
15s interarrival			120s interarrival		
Application	**Bytes**	**Percentage**	**Application**	**Bytes**	**Percentage**
UnKnown	200293593	82.95	UnKnown	196590817	81.42
http	22829451	9.45	http	26120173	10.82
edonkey	5245556	2.17	edonkey	5432314	2.25
pop3	3006653	1.25	pop3	3006653	1.25
msnmessenger	1838597	0.76	msnmessenger	1838790	0.76
smtp	1794988	0.74	smtp	1794988	0.74
dns	1543179	0.64	dns	1543179	0.64
ssl	1139292	0.47	Several	1399913	0.58
bittorrent	1120803	0.46	ssl	1173686	0.49
Several	1064115	0.44	bittorrent	1131694	0.47
skypetoskype	879925	0.36	skypetoskype	800548	0.33
fasttrack	462977	0.19	socks	453792	0.19

socks	388804	0.16	fasttrack	399177	0.17
rtp	207256	0.09	rtp	260078	0.11
xunlei	144369	0.06	nbns	213098	0.09
nbns	128474	0.05	xunlei	144369	0.06
netbios	84805	0.04	stun	137567	0.06
stun	69439	0.03	netbios	84588	0.04
napster	47304	0.02	pplive	63990	0.03
smb	41047	0.02	napster	48376	0.02
soulseek	39233	0.02	smb	41264	0.02
yahoo	34084	0.01	soulseek	39233	0.02
pplive	31542	0.01	yahoo	34084	0.01
ftp	15595	0.01	qq	21822	0.01
qq	15262	0.01	rtsp	20347	0.01
x11	13308	0.01	ftp	15595	0.01
rtsp	12093	0.01	armagetron	10167	0
marca	5407	0	freenet	8330	0
armagetron	4335	0	marca	5407	0
abc	2615	0	x11	4507	0
megauploadAcceso	2462	0	abc	2615	0
sip	2238	0	megauploadAcceso	2462	0
megaupload	1952	0	sip	2238	0
nntp	1460	0	megaupload	1952	0
youTubeClick	1460	0	nntp	1460	0
rlogin	1460	0	youTubeClick	1460	0
elMundo	1278	0	rlogin	1460	0
freenet	1272	0	elMundo	1278	0
gnutella	1230	0	gnutella	1230	0
ntp	516	0	ntp	516	0
elPais	275	0	elPais	275	0
irc	102	0	irc	102	0
imap	51	0	imap	61	0
lpd	12	0	lpd	12	0
TOTAL	**241455754**	**100**	**TOTAL**	**241455754**	**100**

Tabla 4-15: Porcentaje por Aplicaciones Muestreado 1%

Se muestra ahora la distribución del volumen de tráfico identificado por aplicaciones agregado según los grupos ya especificados en el capítulo tercero, para ambos tiempos máximos entre llegadas.

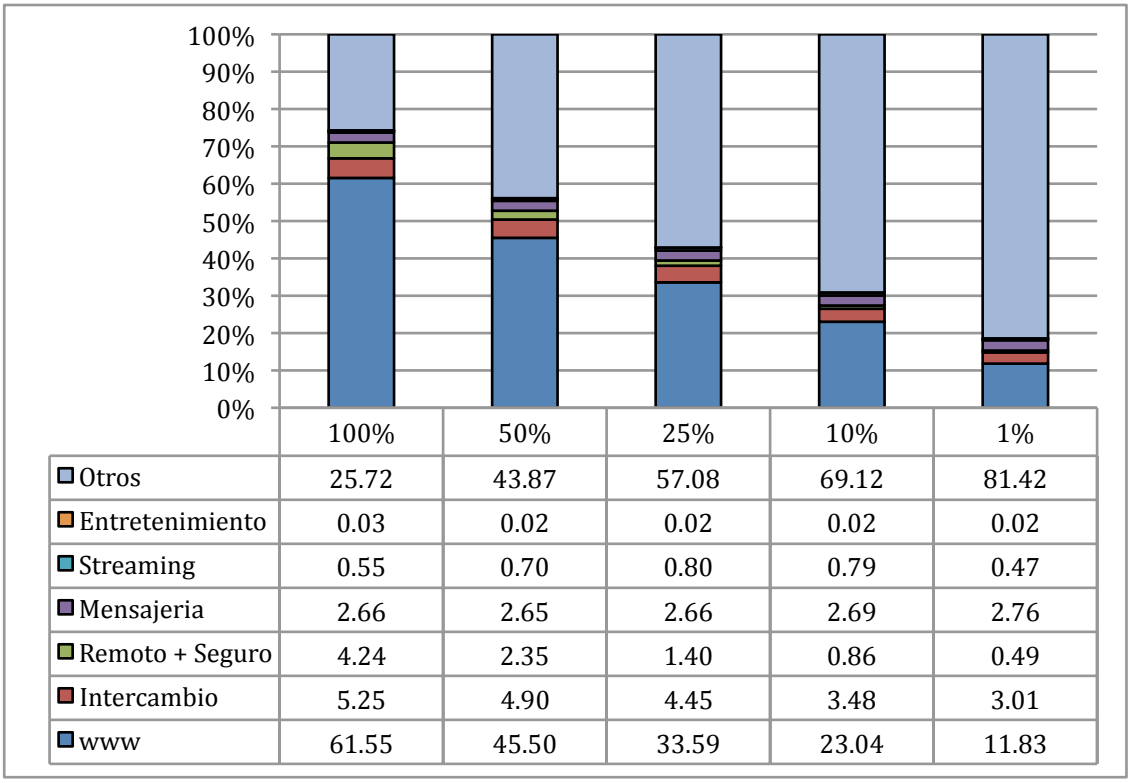
15 segundos interarrival:



	100%	50%	25%	10%	1%
Otros	31.56	50.90	64.76	75.36	82.95
Entretenimiento	0.03	0.02	0.02	0.02	0.02
Streaming	0.67	0.74	0.76	0.67	0.47
Mensajeria	2.66	2.65	2.66	2.66	2.76
Remoto + Seguro	3.66	1.97	1.09	0.68	0.48
Intercambio	3.69	3.04	2.87	2.89	2.95
www	57.74	40.68	27.85	17.72	10.37

Ilustración 4-15: Porcentaje de tráfico por aplicaciones. 15s interarrival. Muestreado.

120 segundos interarrival:



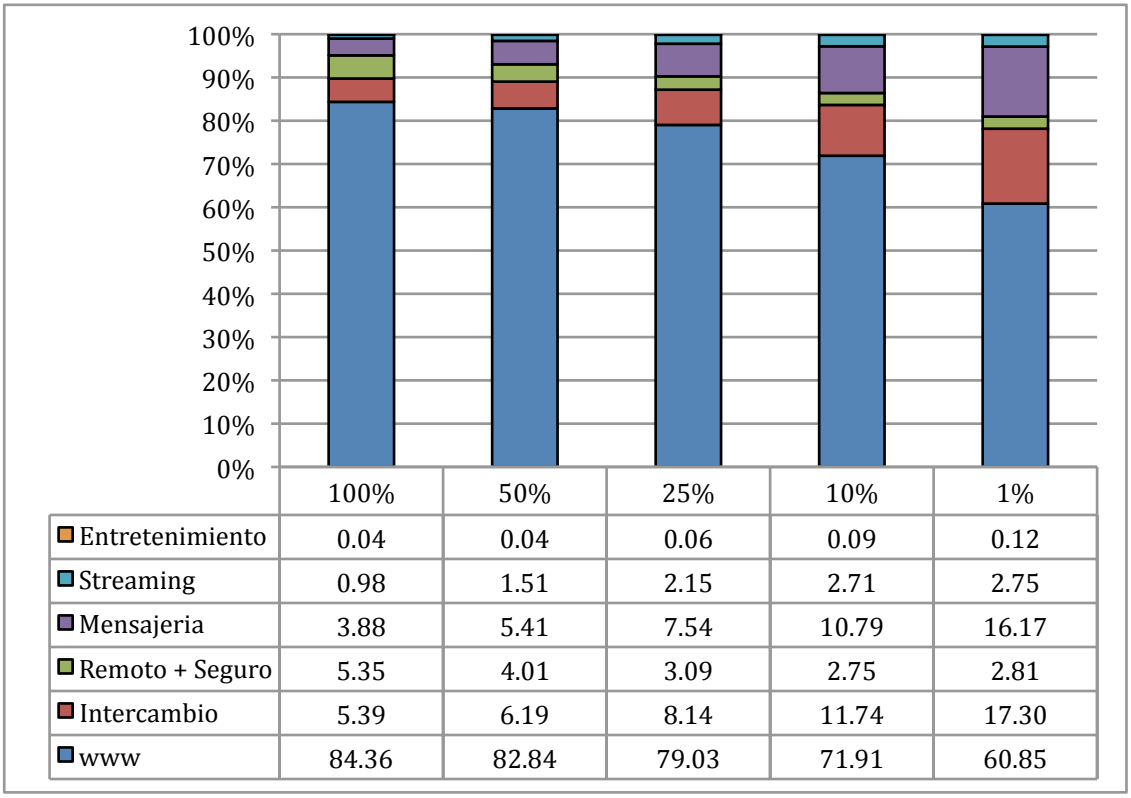
	100%	50%	25%	10%	1%
Otros	25.72	43.87	57.08	69.12	81.42
Entretenimiento	0.03	0.02	0.02	0.02	0.02
Streaming	0.55	0.70	0.80	0.79	0.47
Mensajeria	2.66	2.65	2.66	2.69	2.76
Remoto + Seguro	4.24	2.35	1.40	0.86	0.49
Intercambio	5.25	4.90	4.45	3.48	3.01
www	61.55	45.50	33.59	23.04	11.83

Ilustración 4-16: Porcentaje de tráfico por aplicaciones. 120s interarrival. Muestreado.

Ambos gráficos evidencian el impacto del muestreo en la capacidad de identificación de aplicaciones, que se ve claramente mermada para ambos casos, pasando el porcentaje de flujos no identificados de un 31,56% en la traza original para 15 segundos interarrival y 25.72% para 120 segundos, a un 82.95% para 15 segundos y 81,42% para 15 segundos con una tasa de muestreo de 1 de cada 100.

Para tener una visión más clara de la evolución del porcentaje por grupos de aplicaciones identificadas, se muestran a continuación nuevas gráficas con la distribución de aplicaciones, sin tener en cuenta aquellos flujos no identificados:

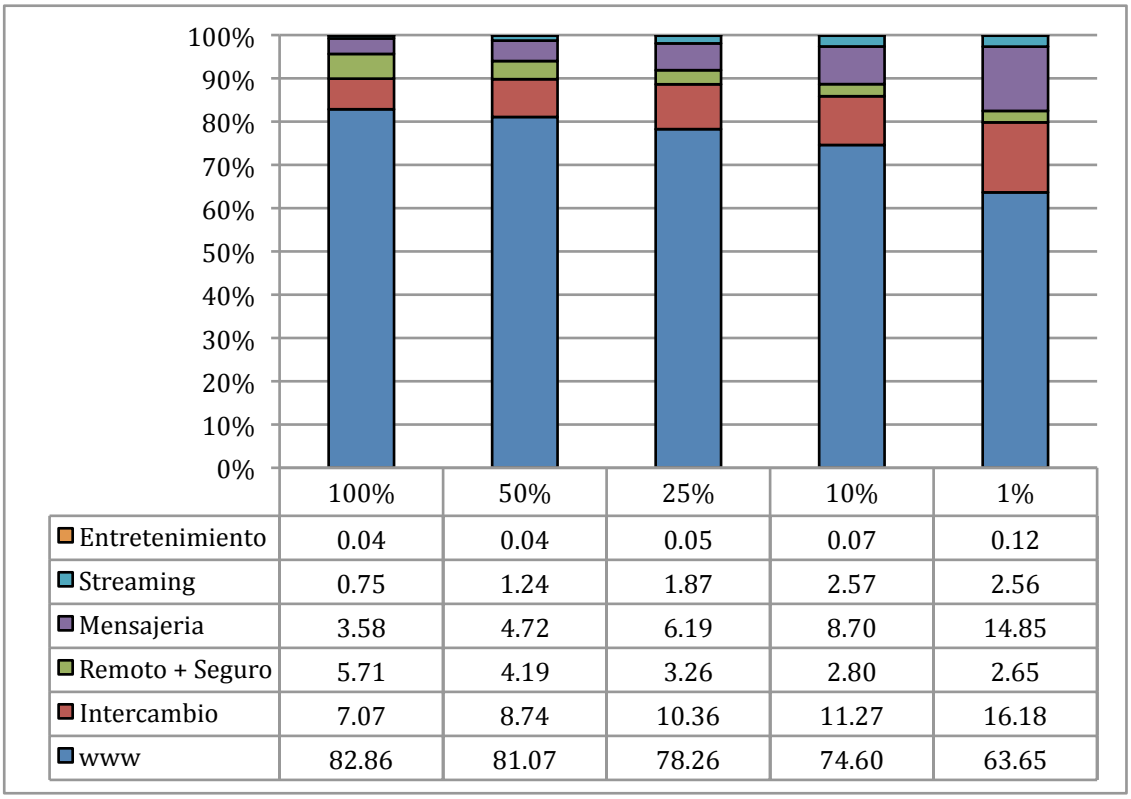
15 segundos interarrival:



	100%	50%	25%	10%	1%
Entretenimiento	0.04	0.04	0.06	0.09	0.12
Streaming	0.98	1.51	2.15	2.71	2.75
Mensajeria	3.88	5.41	7.54	10.79	16.17
Remoto + Seguro	5.35	4.01	3.09	2.75	2.81
Intercambio	5.39	6.19	8.14	11.74	17.30
www	84.36	82.84	79.03	71.91	60.85

Ilustración 4-17: Porcentaje de tráfico identificado. 15s interarrival. Muestreado.

120 segundos interarrival:



	100%	50%	25%	10%	1%
Entretenimiento	0.04	0.04	0.05	0.07	0.12
Streaming	0.75	1.24	1.87	2.57	2.56
Mensajeria	3.58	4.72	6.19	8.70	14.85
Remoto + Seguro	5.71	4.19	3.26	2.80	2.65
Intercambio	7.07	8.74	10.36	11.27	16.18
www	82.86	81.07	78.26	74.60	63.65

Ilustración 4-18: Porcentaje de tráfico identificado. 120s interarrival. Muestreado.

De estos resultados se puede extraer la tendencia que siguen algunos grupos, que tienden a crecer en porcentaje, en detrimento de *WWW* y *Remoto + Seguro*. Este efecto podría deberse a una supuesta dificultad de identificación de flujos con acceso Web y SSL, que resultarían ser las aplicaciones más afectadas a efectos de su identificación conforme disminuye la tasa de muestreo de paquetes.

Este hecho, de ser cierto, debería tenerse en cuenta a la hora de adoptar un sistema de monitorización con muestreo de paquetes, y de este efecto podría surgir una línea de trabajo que tratara de disminuir el impacto del mismo para la identificación de flujos con conexiones a páginas Web.

4.2.7 Flash Flows

Se muestran los resultados de porcentaje de número total de flujos detectados que son en efecto *flash flows*, y el tamaño relativo que tienen en total en comparación con el volumen de datos capturado para cada caso.

15 segundos interarrival:

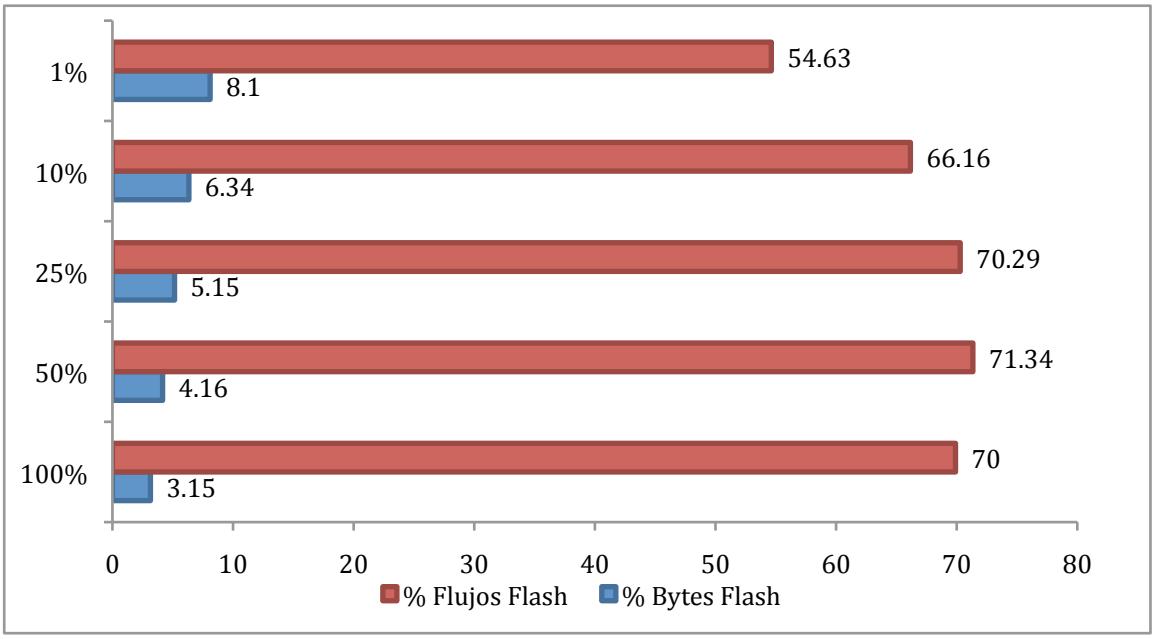


Ilustración 4-19: Flash Flows vs Número de Bytes. 15s Interarrival. Muestreado

120 segundos interarrival:

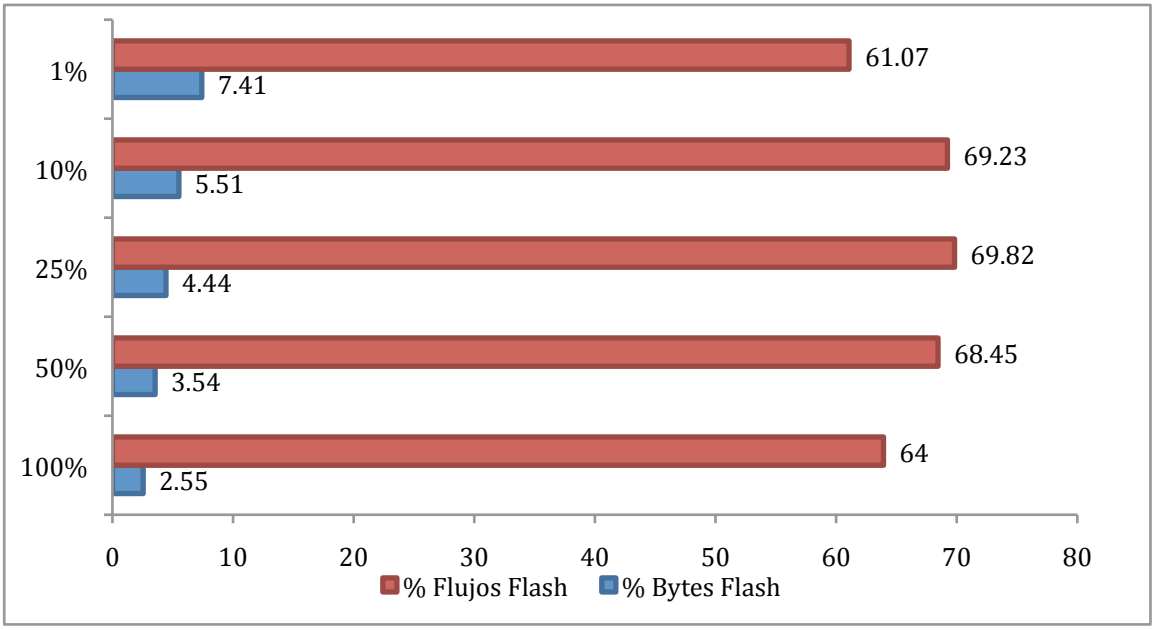


Ilustración 4-20: Flash Flows vs Número de Bytes. 120s Interarrival. Muestreado

Ambas gráficas evidencian la gran diferencia entre el número de flujos que generan los *flash flows* frente al número de bytes de carga que en realidad representan. Resulta interesante, por otro lado, cómo la tendencia creciente que siguen al disminuir la tasa de muestreo, se interrumpe en el caso de un muestreo de 1 de cada 100 paquetes. Esto puede indicar que es a partir de esta tasa, o un valor cercano, cuando comienza un cierto

enmascaramiento del impacto de los *flash flows* en los informes generados. En el caso, entonces de querer adoptar soluciones para la detección exclusiva de grande flujos, como la ya mencionada en [19], éste podría en efecto ser un nivel de muestreo a considerar. En cualquier caso estos resultados muestran la gran importancia que tienen este tipo de tráfico (en número de flujos) y motiva su futuro estudio más detallado.

4.3 Muestreo Distribuido

La propuesta del concepto del muestreo distribuido nace a raíz del efecto que tiene en los sistemas de monitorización mediante muestreo convencional la creación de informes con información en ocasión redundante, que genera grandes cantidades de datos, poco útiles, y que no proporcionan una cobertura amplia del caudal de datos que en realidad atraviesa la red monitorizada.

Un sistema que haga uso de es mismo muestreo y que a la vez sea capaz de aligerar el peso de la captura de paquetes entre nodos y que proporcione aún una mejor cobertura del total de datos que pasan por la red, es un concepto que podría aportar muchos beneficios al operador de la misma.

Para lograr una captura de paquetes entre equipos que evite el solapamiento de datos se hace uso del concepto de *offset*. Desde el punto de vista de la implementación en el generador de informes *netflow* utilizado en este estudio, este *offset* representa la posición del primer paquete a muestrear dentro de la traza de datos, que se abre en modo *offline*. Utilizando distintos valores de *offset* para ejecuciones con las misma tasa de muestreo se garantiza que los paquetes capturados en cada caso son siempre diferentes, y por lo tanto una agregación de los datos obtenidos por cada una proporcionará información más amplia y fidedigna del tráfico monitorizado que si se tomaran informes por separado, sin coordinación alguna.

La implantación de un sistema así en una red con un sistema de monitorización que actúe en tiempo real podría implicar la necesidad de algún tipo de comunicación entre nodos para evitar la captura de la misma secuencia de paquetes. Esto se podría lograr de manera relativamente sencilla mediante la inclusión de una bandera en paquetes de referencia, que se utilizarían para aplicar el *offset* en la captura entre nodos ya mencionado. Los resultados mostrados a continuación, sin embargo, tratan de comprobar en qué medida es posible no

realizar esa coordinación entre *offsets* y cual es su efecto en los informes generados observando los resultados para diversas combinaciones, y comprobando su variabilidad.

Se plantea, por lo tanto, una posible solución basada en el sistema de muestreo coordinado de flujos expuesto en [16]. Este sistema hace uso de una asignación de rangos dentro de el espacio de valores posibles de la función *hash* adoptada para la identificación de flujos. De este modo, se otorga un intervalo, o *manifiesto*, a cada equipo o equipos de la red, de manera que se garantiza que cada equipo sólo capturará una determinada porción de tráfico concreto. Según el estudio, la cobertura alcanzada en términos de extensión de captura de los datos que atraviesan la red es muy buena y el sistema es robusto a cambios temporales en tendencias de tráfico. Existe , sin embargo, la limitación de que, para cada paquete que pasa por el nodo de identificación de flujos, es necesario hacer una lectura de las cabeceras del mismo para la computación del código *hash* que es entonces comparado con el *manifiesto* otorgado, a partir del cual se realizaría un análisis del paquete y su inclusión en la tabla de flujos local. Esto hace necesario el uso de memorias SRAM lo suficientemente rápidas para la lectura de todos los paquetes que pasen por el equipo a velocidad de línea. Si bien el estudio hace una aproximación sobre el tamaño requerido, y consideran que este es relativamente pequeño en cuanto que la cobertura del trafico capturado sería mayor con la misma cantidad de SRAM que para un sistema convencional, el uso de un sistema de muestreo distribuido de paquetes añadido a este de distribución de tráfico por porciones puede aportar la disminución en la velocidad de lectura necesaria para evitar el uso de estas memorias, o de reducir significativamente su tamaño requerido.

Se muestra un esquema de funcionamiento del sistema planteado en la figura 4-21 que incluye un muestreo coordinado de flujos combinado con un muestro distribuido de paquetes entre equipos con el mismo *manifiesto*:

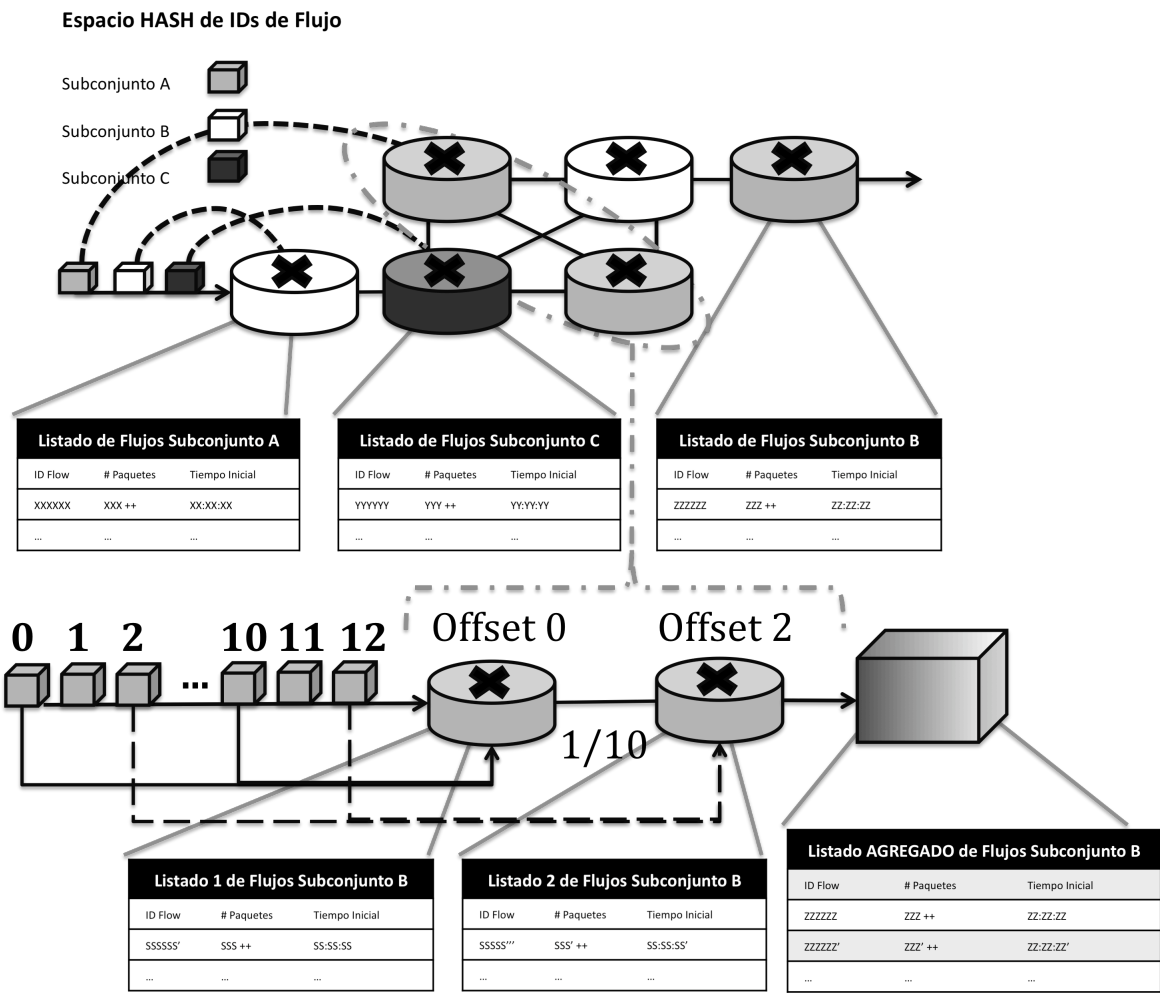


Ilustración 4-21: Esquema de funcionamiento de muestreo distribuido de paquetes sobre sistema de monitorización basado en asignación de flujos.

La figura 4-21 muestra una red de monitorización con nodos generadores de flujos que tienen asignados rangos concretos del tráfico total. Estos subconjuntos, que son identificados mediante una función hash se muestran aquí según distintas tonalidades, tanto en los paquetes de entrada a la red, como en los propios nodos. Según este esquema, por lo tanto, los paquetes pertenecientes al rango o subconjunto A del espacio hash sólo serán muestreados por los generadores que tienen asignado el subconjuto A en su *manifiesto*.

El esquema posteriormente se extrae de esta asignación de rangos y muestra cual es el funcionamiento de uno de esto subconjuntos, que en este caso estaría siendo monitorizado por dos equipos. Estos dos equipos reciben un caudal de datos que se asume por sencillez que es el mismo. Si los dos realizan, dentro de su rango asignado, un subsecuente muestreo con tasa 1 de cada 10, pero con distinto valor de *offset*, 0 en un caso y 2 en otro, ambos nodos analizarían porciones distintas, evitando la duplicidad en sus informes. Estos

informes locales deberan ser finalmente enviados a un colector que se encargue de hacer una agregación de los mismos, de manera que ambos se completen, logrando presumiblemente así una tasa real de muestreo de tráfico mayor al 10% que cada uno de los generadores aplica individualmente.

Para poder hacer una valoración de la eficiencia del sistema propuesto, se plantea un caso concreto a partir del cual se realizarán distintas ejecuciones y se estudiarán los resultados de las métricas ya consideradas en apartados anteriores.

Según este esquema existen, dentro de una red de datos con un sistema de monitorización basado en muestreo de flujo, dos nodos con un mismo *manifiesto* asignado, de manera que se asigna a estos el muestreo de un rango concreto del trafico total que pasa por la red. Basándonos en los resultados de cobertura ya mencionados en [16], se asume que esta es total en éste caso, y por lo tanto, todos los paquetes que pasan por uno de los nodos, pasan en efecto también por el otro. Se plantea por tanto la distribución del muestreo de una misma traza de datos, que se efectuará de manera coordinada a través de un valor de *offset*. En caso de no coordinación en el *offset* de los distintos generadores de flujos podría darse el caso de que dos nodos que generan informes netflow analizaran exactamente el mismo subconjunto de paquetes. En este caso, estaríamosen una situación similar a la presentada en la sección anterior. En el caso de offset distintos, la agregación de los datos obtenidos por cada uno de los nodos, proporciona en principio unos resultados mejores que aquellos obtenidos por muestreo simple. La siguiente sección trata de evaluar estos resultados y de comprobar cómo la variabilidad en el uso de distintas combinaciones de *offset* entre nodos puede afectar a los mismos.

Para ello se realiza una nueva modificación al generador de flujos implementado, de manera que este agregue automáticamente los resultados obtenidos con dos *offets* distintos a la vez. Se efectúa a continuación una nueva batería de ejecuciones para distintas tasas de muestreo y *offsets*, que tratan de cubrir todas las diferencias posibles entre los mismos. En esta ocasión se decide limitar los resultados mostrados a 15 segundos como único tiempo máximo entre llegadas. Estas son las ejecuciones planteadas:

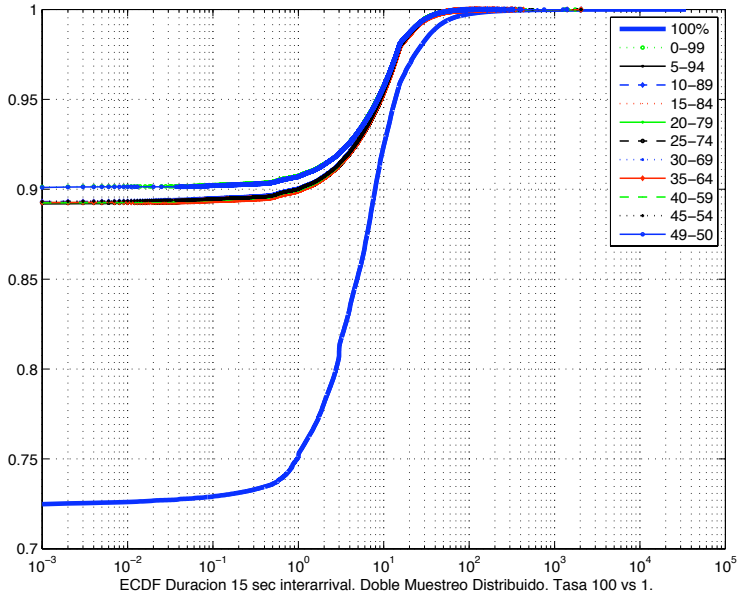
- Doble muestreo 1 de cada 4. Tasa total equivalente 50%. 6 ejecuciones.
    - Offset1 = 0; Offset2 = 1
    - Offset1 = 0; Offset2 = 2
    - Offset1 = 0; Offset2 = 3
    - Offset1 = 1; Offset2 = 2
    - Offset1 = 1; Offset2 = 3
    - Offset1 = 2; Offset2 = 3
- Doble muestreo 1 de cada 10. Tasa total equivalente 20%. 9 ejecuciones.
    - Offset1 = 0; Offset2 = 8
    - Offset1 = 0; Offset2 = 9
    - Offset1 = 1; Offset2 = 7
    - Offset1 = 1; Offset2 = 8
    - Offset1 = 2; Offset2 = 6
    - Offset1 = 2; Offset2 = 7
    - Offset1 = 3; Offset2 = 5
    - Offset1 = 3; Offset2 = 6
    - Offset1 = 4; Offset2 = 5
- Doble muestreo 1 de cada 10. Tasa total equivalente 20%. 11 ejecuciones.
    - Offset1 = 0; Offset2 = 99
    - Offset1 = 5; Offset2 = 94
    - Offset1 = 10; Offset2 = 89
    - Offset1 = 15; Offset2 = 84
    - Offset1 = 20; Offset2 = 79
    - Offset1 = 25; Offset2 = 74
    - Offset1 = 30; Offset2 = 69
    - Offset1 = 35; Offset2 = 64
    - Offset1 = 40; Offset2 = 59
    - Offset1 = 45; Offset2 = 54
    - Offset1 = 49; Offset2 = 50

A continuación se muestran los resultado obtenidos con estas ejecuciones para cada métrica a considerar

4.3.1 Duración

Junto con cada grafica se incluyen los valores *Chi Cuadrado* y *Phi* obtenidos de la ejecución de los tests de similitud. Por cada tasa se muestra la grafica de distribución acumulada con distintos valores de *offset*, que se marcan con patrones diferentes de color, punteado y símbolos, frente a los resultados del muestreo al 100%, siempre en color azul. Al utilizar distintos *offsets*, se puede observar directamente cómo puede llegar a variar el resultado según este parámetro.

Tasa 2x1%

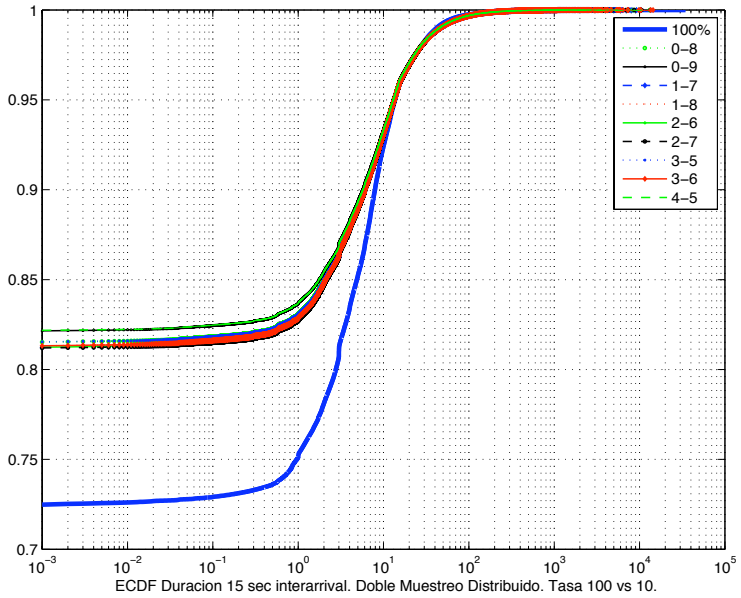


15 interarrival.
Tasa = 2x1%.
22 G Libtad
Chi Ref = 33.9244
Alfa = 0.95

Offsets	Chi Test	Phi
0-99	1.3750×10^4	0.0232
5-94	1.3585×10^4	0.0232
10-89	1.3750×10^4	0.0233
15-84	1.3828×10^4	0.0234
20-79	1.3820×10^4	0.0234
25-74	1.3672×10^4	0.0232
30-69	1.3554×10^4	0.0231
35-64	1.3750×10^4	0.0233
40-59	1.3554×10^4	0.0231
45-54	1.3789×10^4	0.0233
49-50	1.3567×10^4	0.0232

Ilustración 4-22: CDF Duración Muestreo Distribuido 2x1%. 15s Interarrival

Tasa 10%

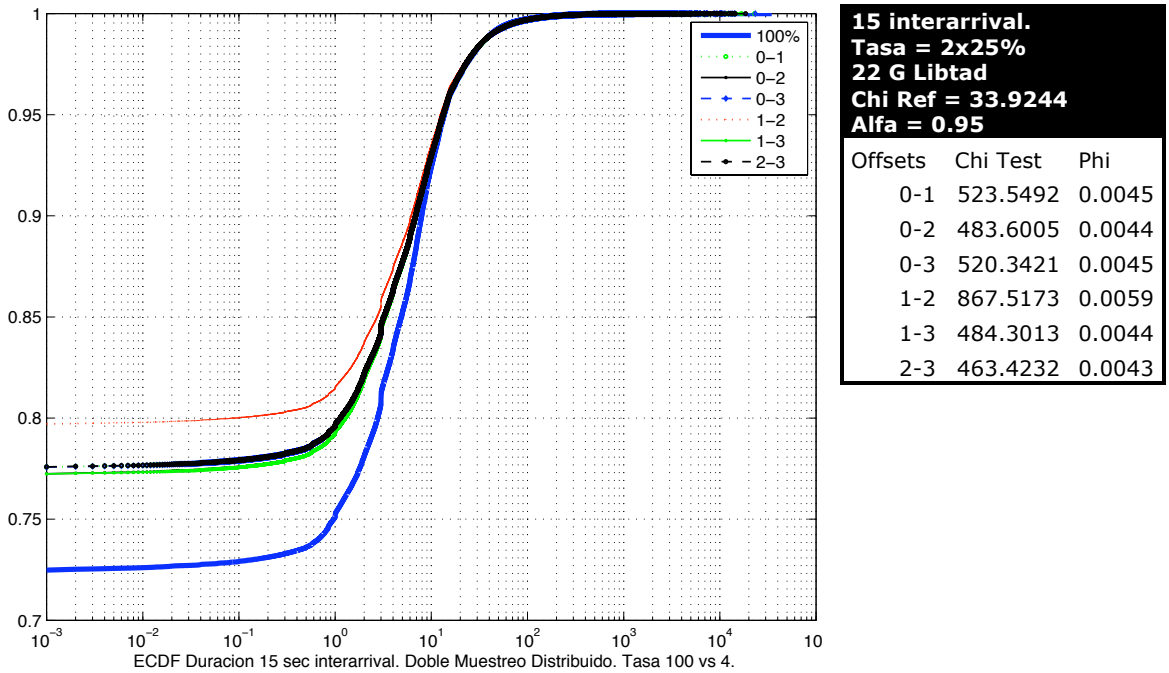


15 interarrival.
Tasa = 2x10%
22 G Libtad Chi Ref = 33.9244
Alfa = 0.95

Offsets	Chi Test	Phi
0-8	822.0823	0.0057
0-9	648.5623	0.0051
1-7	925.0089	0.0061
1-8	921.2892	0.0060
2-6	958.0021	0.0062
2-7	869.1458	0.0059
3-5	909.7885	0.0060
3-6	935.6966	0.0061
4-5	747.3246	0.0054

Ilustración 4-23: CDF Duración Muestreo Distribuido 2x10%. 15s Interarrival

Tasa 25%



15 interarrival.
Tasa = 2x25%
22 G Libtad
Chi Ref = 33.9244
Alfa = 0.95

Offsets	Chi Test	Phi
0-1	523.5492	0.0045
0-2	483.6005	0.0044
0-3	520.3421	0.0045
1-2	867.5173	0.0059
1-3	484.3013	0.0044
2-3	463.4232	0.0043

Ilustración 4-24: CDF Duración Muestreo Distribuido 2x25%. 15s Interarrival

De las gráficas obtenidas, y de su comparación con los resultados obtenidos con muestreo simple, se observa cómo en efecto se produce la mejora esperada en las distribuciones y en los valores *de Chi Cuadrado* y *Phi*. Se recuerda que, en el caso de la duración de flujo, no se realiza ninguna transformación de los valores obtenidos para hacer la comparación con la gráfica de la traza completa.

Al ser ambos valores de los test de similitud calculados a través de ecuaciones no lineales, no podemos asumir una relación de linealidad en la mejoría de los resultados para muestreo simple y doble distribuido, pero si se puede constatar cómo en efecto esta mejoría se produce cuantitativamente. Por ejemplo, para una tasa de muestreo de 10% para el caso simple, el valor de *Phi* obtenido es de 0.0072, mientras que para muestreo doble al 10%, que debería corresponder aproximadamente con un 20% simple, este valor oscila entre 0.0051 y 0.0062. Este resultado apunta a que un muestreo doble que no tenga una coordinación entre los *offsets* de cada nodo, pero que agregue los informe *netflow* obtenidos por cada uno obtiene siempre resultados al menos tan buenos como los de muestreo simple a esa tasa, pero que en la mayoría de los casos llegarán a ser significativamente mejores si los *offsets* no son exactamente el mismo.

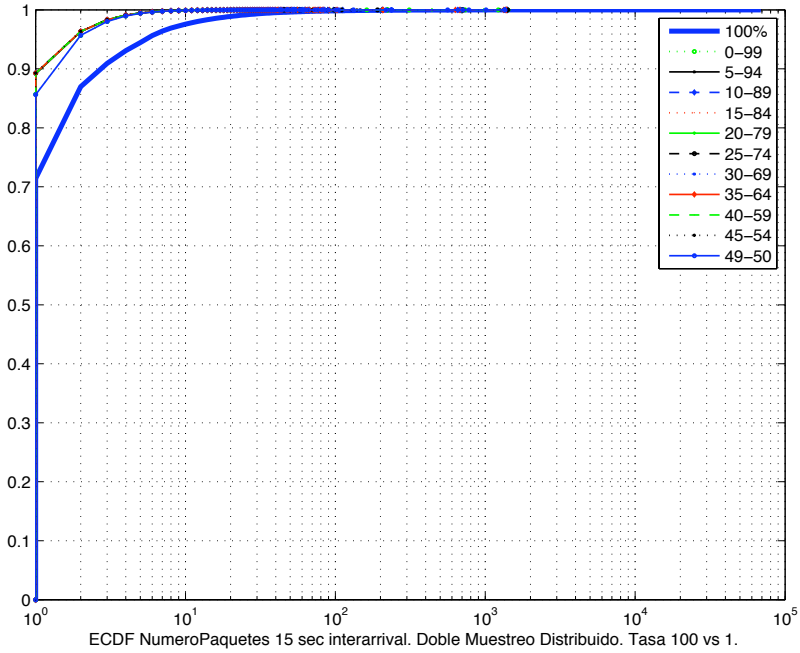
Resulta en este caso de especial interés el caso de muestreo doble al 25%, que correspondería con uno simple de 50%. En este caso, se cuenta con resultados de muestreo simple correspondientes a esa tasa, y por lo tanto se puede comprobar en qué medida es posible alcanzarlos usando la mitad de tasa por cada nodo y agregando sus informes.

Para muestreo simple al 50%, el valor de *Phi* es 0.0044, mientras que para doble al 25%, estos son en su mayoría iguales o incluso ligeramente mejores, a excepción de una única combinación de *offsets*, en este caso 1-2, que tiene resultados peores. De 6 combinaciones de offset *posibles*, 5 son iguales o mejores y 1 es peor.

Es interesante apuntar cómo, en el caso de las otras dos tasas, 2x10% y 2x1%, siempre existen también dos posible "curvas" en las gráficas de distribución, una de las cuales es siempre minoritaria en su aparición. Estas curvas minoritarias, sin embargo, no parecen tener siempre resultados peores que la media en los tests de similitud. Existe, por otro lado, una característica común entre ellas para cada tasa, y es que siempre corresponden a *offsets* que tienen sólo una unidad de diferencia. En el caso de 2x1%, estas corresponden a *offsets* de 0-99 y 49-50, mientras que 2x10%, corresponden a 0-9 y 4-5. Esta fenómeno indica la aparición de algún efecto diferente que se produce de manera específica para muestreo de valores de *offset* consecutivos, si bien de momento no se puede asumir que se traduzcan en resultados siempre peores que el resto de combinaciones.

4.3.2 Número de Paquetes

Tasa 1%

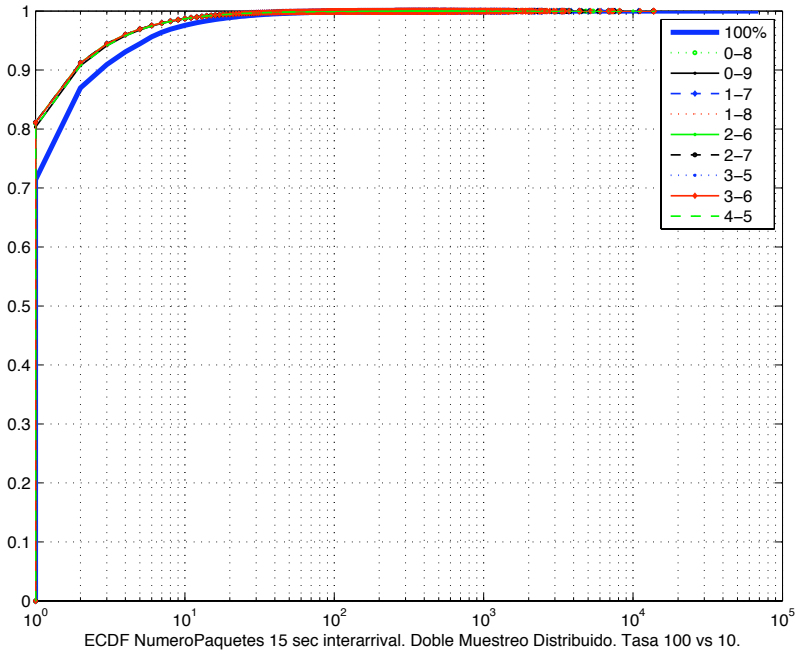


15 interarrival.
Tasa = 2x1%
19 G Libtad
Chi Ref = 30.1435
Alfa = 0.95

Offsets	Chi Test	Phi
0-99	4.7268×10^5	0.1367
5-94	3.5234×10^5	0.1180
10-89	3.5017×10^5	0.1177
15-84	3.7084×10^5	0.1211
20-79	3.9865×10^5	0.1255
25-74	3.6723×10^5	0.1205
30-69	3.8170×10^5	0.1228
35-64	3.8709×10^5	0.1237
40-59	3.8308×10^5	0.1231
45-54	3.7187×10^5	0.1212
49-50	4.8303×10^5	0.1382

Ilustración 4-25: CDF Número Paquetes Muestreo Distribuido 2x1%. 15s Interarrival

Tasa 10%

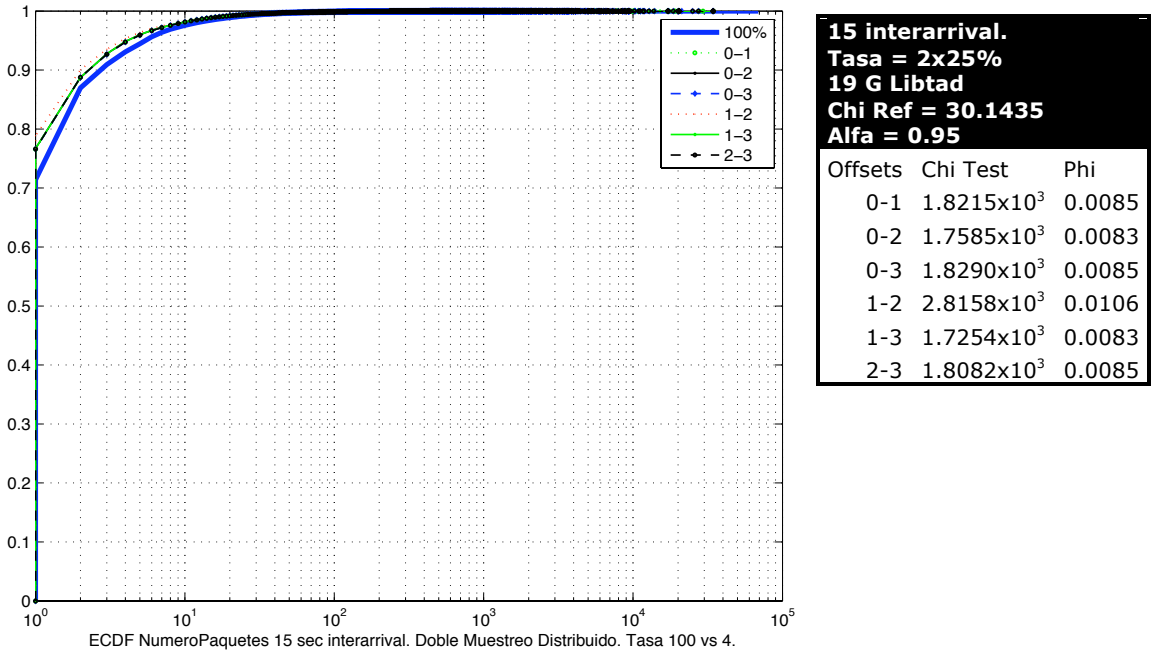


15 interarrival.
Tasa = 2x10%
19 G Libtad
Chi Ref = 30.1435
Alfa = 0.95

Offsets	Chi Test	Phi
0-8	1.6642×10^4	0.0256
0-9	1.6869×10^4	0.0258
1-7	1.7227×10^4	0.0261
1-8	1.6977×10^4	0.0259
2-6	1.7172×10^4	0.0261
2-7	1.5834×10^4	0.0250
3-5	1.7571×10^4	0.0264
3-6	1.7114×10^4	0.0260
4-5	1.7191×10^4	0.0261

Ilustración 4-26: CDF Número Paquetes Muestreo Distribuido 2x10%. 15s Interarrival

Tasa 25%



15 interarrival.
Tasa = 2x25%
19 G Libtad
Chi Ref = 30.1435
Alfa = 0.95

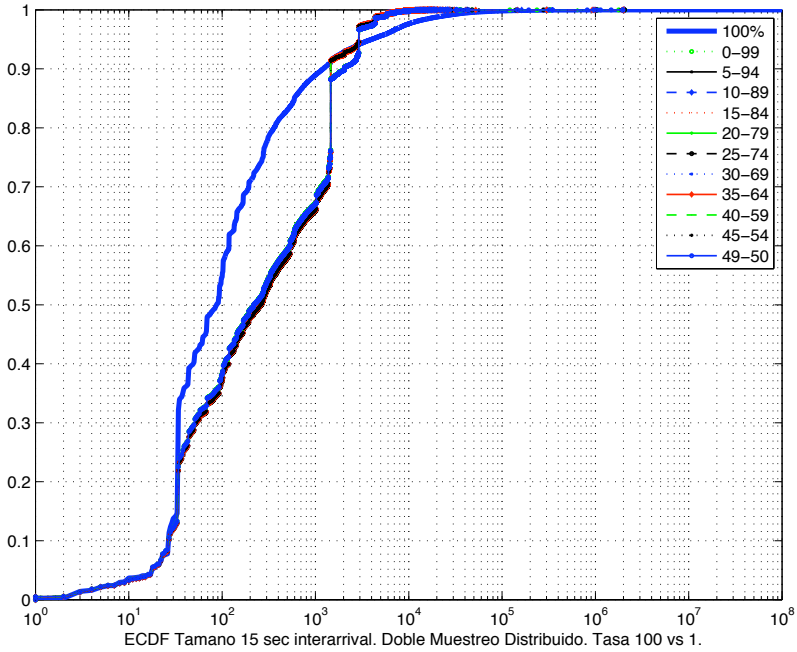
Offsets	Chi Test	Phi
0-1	1.8215×10^3	0.0085
0-2	1.7585×10^3	0.0083
0-3	1.8290×10^3	0.0085
1-2	2.8158×10^3	0.0106
1-3	1.7254×10^3	0.0083
2-3	1.8082×10^3	0.0085

Ilustración 4-27: CDF Número Paquetes Muestreo Distribuido 2x25%. 15s Interarrival

Los resultado extraídos para esta métrica vuelven a confirmar la mejoría esperada en cada uno de los casos, que de nuevo superan a aquellos obtenidos con la misma tasa a muestreo simple. Los resultados de tests de similitud comparados de muestreo doble al 25% no alcanzan el mismo nivel de bondad frente a simple al 50%, aunque si se acercan considerablemente. Se confirma de nuevo la tendencia en las distribuciones de la existencia de una curva minoritaria que se corresponde con combinaciones de *offset* consecutivos. En el caso de muestreo doble al 10% este efecto es poco apreciable, pero en 2x1% y 2x25% corresponden inequívocamente con estos valores de *offset*, particularmente 0-99 y 49-50, y 1-2 respectivamente. En esta ocasión con resultados peores en los test de similitud frente a la media. Resultados que, sin embargo, se siguen manteniendo en cualquier caso por encima de los de muestreo simple.

4.3.3 Tamaño

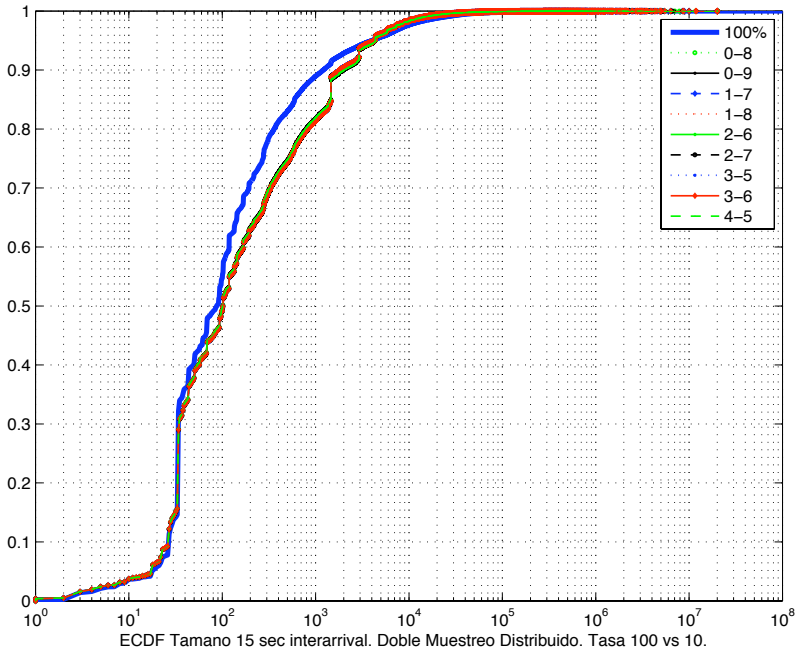
Tasa 1%



15 interarrival. Tasa = 2x1% 15 G Libtad Chi Ref = 24.9958 Alfa = 0.95		
Offsets	Chi Test	Phi
0-99	1.1795×10^5	0.0683
5-94	0.8150×10^5	0.0568
10-89	0.9945×10^5	0.0627
15-84	0.8989×10^5	0.0596
20-79	0.9880×10^5	0.0625
25-74	0.8649×10^5	0.0585
30-69	0.9749×10^5	0.0621
35-64	0.9515×10^5	0.0613
40-59	1.0603×10^5	0.0647
45-54	0.8763×10^5	0.0589
49-50	1.2679×10^5	0.0708

Ilustración 4-28: CDF Tamaño Muestreo Distribuido 2x1%. 15s Interarrival

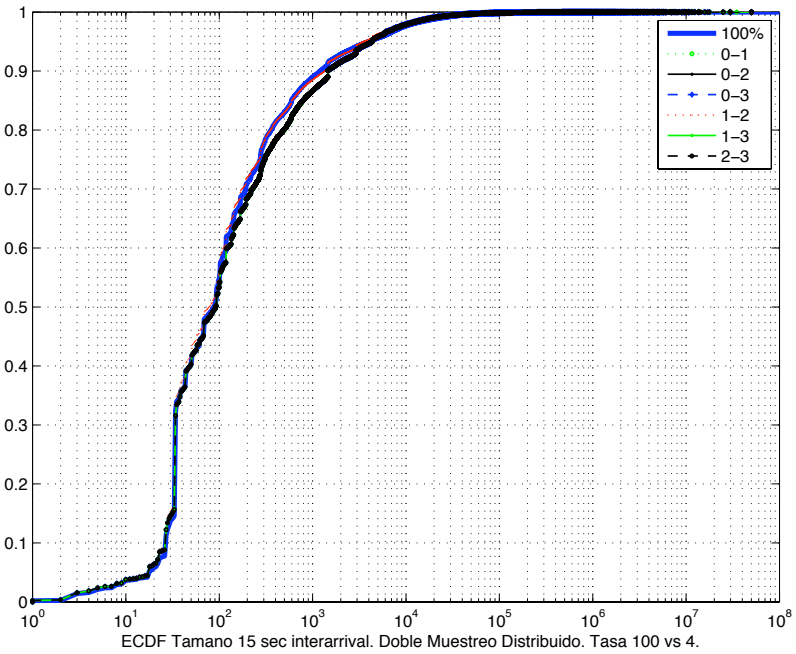
Tasa 10%



15 interarrival. Tasa = 2x10% 15 G Libtad Chi Ref 24.9958 Alfa = 0.95		
Offsets	Chi Test	Phi
0-8	1.0629×10^4	0.0205
0-9	1.0158×10^4	0.0200
1-7	1.0442×10^4	0.0203
1-8	1.0377×10^4	0.0203
2-6	1.0675×10^4	0.0205
2-7	0.9797×10^4	0.0197
3-5	1.0956×10^4	0.0208
3-6	1.0534×10^4	0.0204
4-5	1.0469×10^4	0.0203

Ilustración 4-29: CDF Tamaño Muestreo Distribuido 2x10%. 15s Interarrival

Tasa 25%



15 interarrival. Tasa = 2x25% 15 G Libtad Chi Ref 24.9958 Alfa = 0.95		
Offsets	Chi Test	Phi
0-1	1.1969×10^3	0.0069
0-2	1.1473×10^3	0.0067
0-3	1.2184×10^3	0.0069
1-2	1.9472×10^3	0.0088
1-3	1.1491×10^3	0.0067
2-3	1.1959×10^3	0.0069

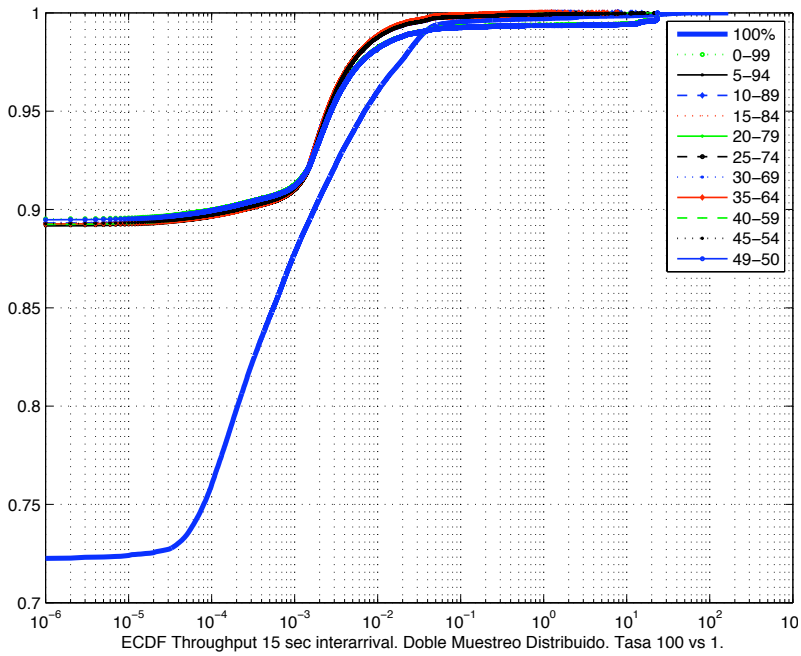
Ilustración 4-30: CDF Tamaño Muestreo Distribuido 2x25%. 15s Interarrival

Los tamaños obtenidos en las ejecuciones con doble muestreo muestran de nuevo distribuciones que siguen la misma tendencia que para muestreo simple en cuanto a la aparición de modas conforme se disminuye la tasa de muestreo. Estas modas siguen estando entorno a 34 y 1460 bytes.

Los valores de las pruebas de similitud con la traza completa evidencian la mejoría que se obtiene en todos los casos mediante la agregación de informes de dos nodos a la misma tasa. Existen también curvas minoritarias en las combinaciones de offset consecutivas ya identificadas para las métricas anteriores, que corresponden con resultados peores que la media.

4.3.4 Throughput

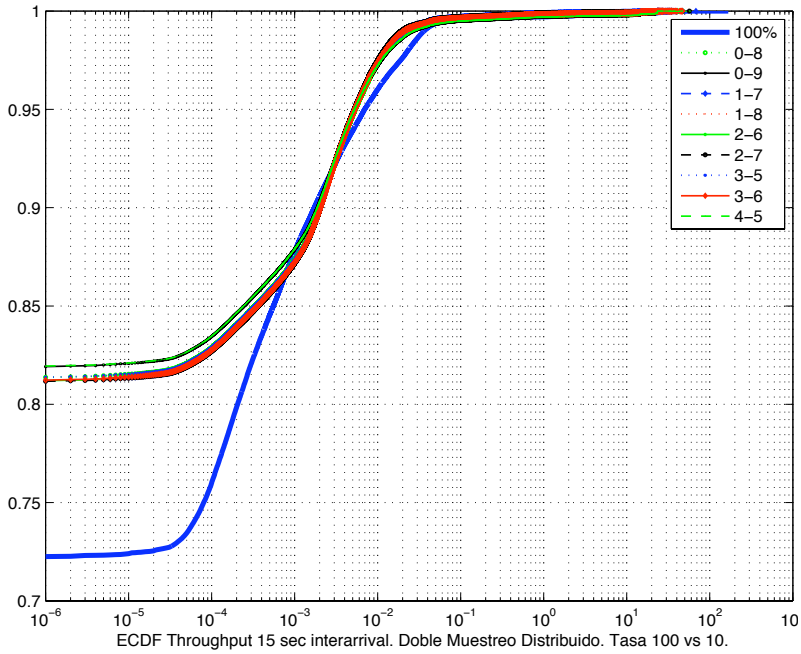
Tasa 1%



15 interarrival. Tasa = 2x1%. 64 G Libtad Chi Ref = 83.6753 Alfa = 0.95		
Offsets	Chi Test	Phi
0-99	1.5775×10^8	2.4973
5-94	0.0096×10^8	0.1946
10-89	0.0033×10^8	0.1146
15-84	0.0031×10^8	0.1113
20-79	0.0032×10^8	0.1121
25-74	0.0031×10^8	0.1115
30-69	0.0031×10^8	0.1101
35-64	0.0031×10^8	0.1103
40-59	0.0035×10^8	0.1180
45-54	0.0173×10^8	0.2616
49-50	1.6891×10^8	2.5841

Ilustración 4-31:CDF Throughput Muestreo Distribuido 2x1%. 15s Interarrival

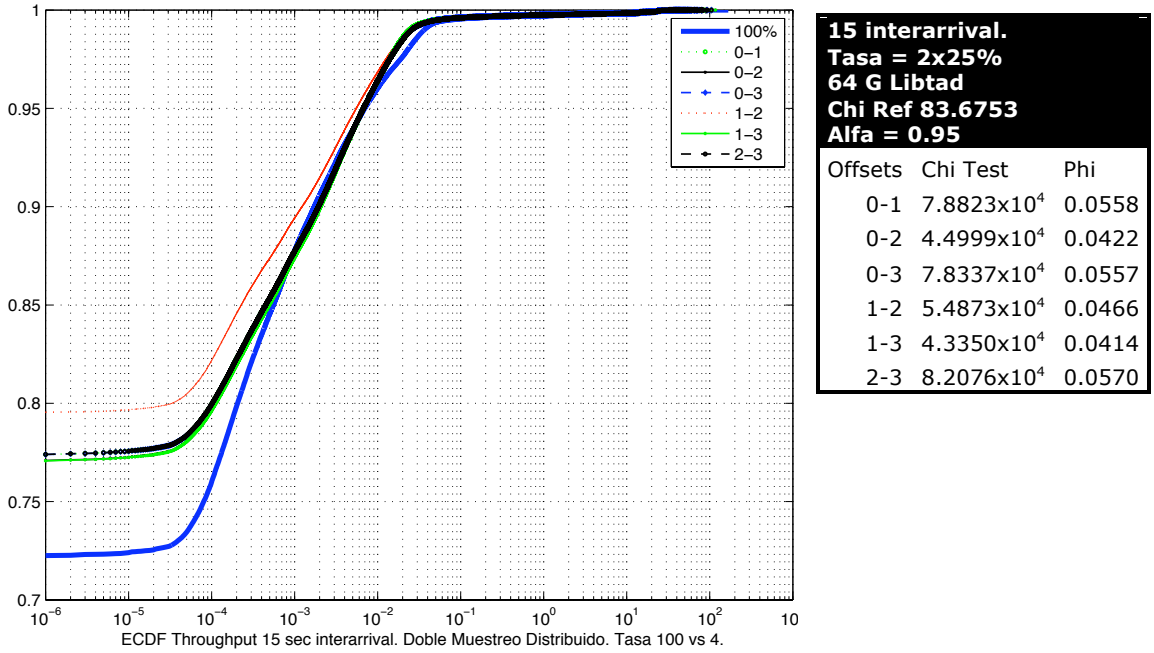
Tasa 10%



15 interarrival. Tasa = 2x10% 64 G Libtad Chi Ref = 83.6753 Alfa = 0.95		
Offsets	Chi Test	Phi
0-8	3.2086×10^6	0.3562
0-9	7.7244×10^4	0.5526
1-7	0.3888×10^4	0.1240
1-8	1.7409×10^4	0.2623
2-6	0.3894×10^4	0.1241
2-7	0.1788×10^4	0.0841
3-5	3.3034×10^4	0.3614
3-6	1.7772×10^4	0.2651
4-5	7.8085×10^4	0.5556

Ilustración 4-32: CDF Throughput Muestreo Distribuido 2x10%. 15s Interarrival

Tasa 25%



15 interarrival.
Tasa = 2x25%
64 G Libtad
Chi Ref 83.6753
Alfa = 0.95

Offsets	Chi Test	Phi
0-1	7.8823×10^4	0.0558
0-2	4.4999×10^4	0.0422
0-3	7.8337×10^4	0.0557
1-2	5.4873×10^4	0.0466
1-3	4.3350×10^4	0.0414
2-3	8.2076×10^4	0.0570

Ilustración 4-33: CDF Throughput Muestreo Distribuido 2x25%. 15s Interarrival

Las gráficas de Throughput muestran cómo existe un grado de variabilidad en los resultados que ponen de nuevo en tela de juicio la conveniencia de usar una métrica definida de este modo. También se puede observar como las curvas obtenidas llegan a tener formas sensiblemente diferentes a la original, especialmente en lo que respecta a la pendiente de su crecimiento, que en las partes de evolución logarítmica, llegan a ser claramente diferentes a simple vista. Los valores de Chi cuadrado y Phi no son siempre mejores que para muestreo simple. Todo esto lleva a la conclusión de que esta métrica es poco adecuada para ambos tipos de muestreo.

El fenómeno de curvas minoritarias sigue existiendo para combinaciones de *offset* ya identificadas, con resultados de test en su mayoría sensiblemente peores del resto. La variabilidad en el resto de resultados, sin embargo, no permite la toma de conclusiones, aunque estos parecen seguir apuntando a una cierta manifestación concreta para estas ejecuciones.

4.3.5 Número total de flujos

Esta es la evolución que sigue el número de flujos detectados con cada tasa de muestreo doble en comparación con la traza original, acorde con la evolución para muestreo simple.

100%	2x25%	2x10%	2x1%
12647520	7740067	3897671	590506

Tabla 4-16: Número Total de Flujos Muestreo Distribuido

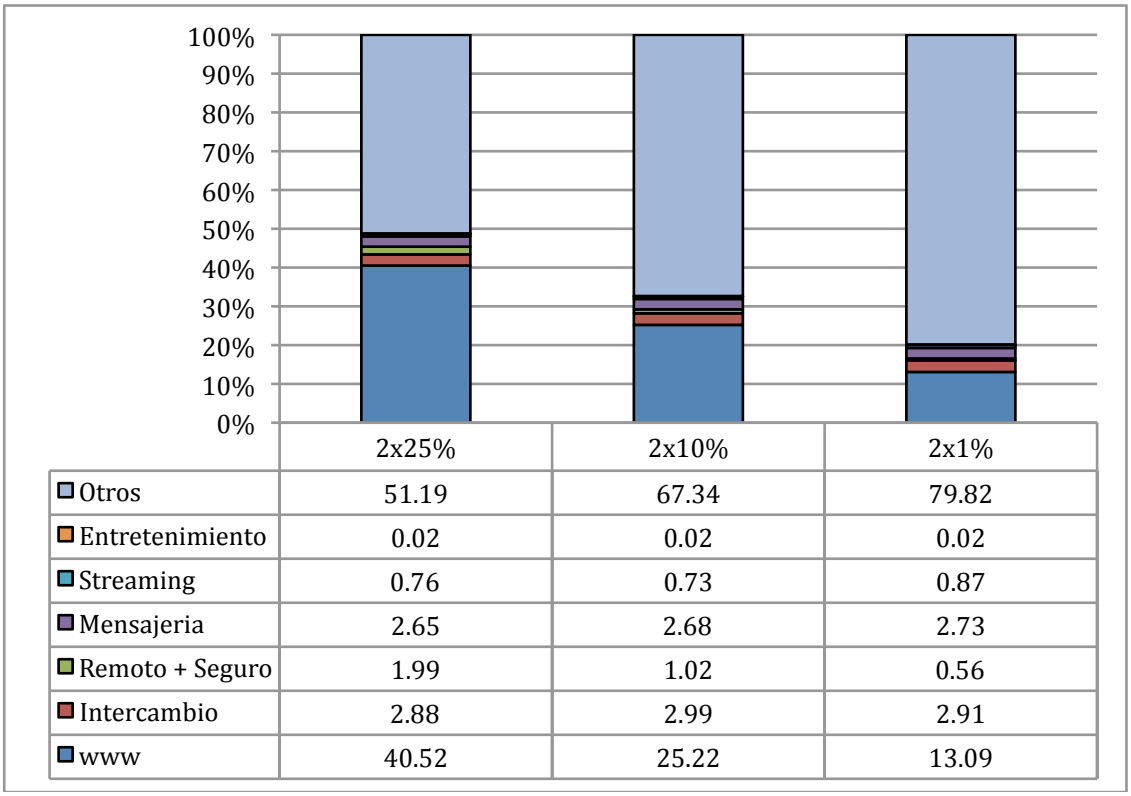
4.3.6 Porcentaje por aplicaciones

2x25%			2x10%			2x1%		
Application	**Bytes**	**%**	**Application**	**Bytes**	**%**	**Application**	**Bytes**	**%**
UnKnown	6172627570	51.19	UnKnown	3247597122	67.34	UnKnown	385300972	79.82
http	4707797565	39.04	http	1135199973	23.54	http	56577805	11.72
edonkey	261007078	2.16	edonkey	109389241	2.27	edonkey	10393466	2.15
ssl	238123937	1.97	pop3	57816818	1.2	pop3	5944839	1.23
pop3	143548522	1.19	ssl	48712281	1.01	msnmessenge	3625242	0.75
msnmessenge	88291568	0.73	msnmessenge	35507064	0.74	smtp	3580299	0.74
smtp	85708502	0.71	smtp	34374317	0.71	dns	3082704	0.64
dns	77194862	0.64	dns	30890676	0.64	ssl	2643718	0.55
socks	61076389	0.51	nbns	27545567	0.57	nbns	2425666	0.5
bittorrent	57794479	0.48	bittorrent	22327324	0.46	rtp	2406330	0.5
rtsp	45074370	0.37	socks	17322677	0.36	bittorrent	2217418	0.46
nbns	28764464	0.24	rtsp	11909610	0.25	skypetoskype	1695656	0.35
skypetoskype	24562010	0.2	skypetoskype	11752711	0.24	fasttrack	860241	0.18
rtp	16242417	0.13	rtp	7233684	0.15	socks	767140	0.16
fasttrack	13486803	0.11	fasttrack	6688046	0.14	xunlei	286702	0.06
stun	9317720	0.08	stun	4532733	0.09	stun	203456	0.04
xunlei	7628462	0.06	pplive	4275720	0.09	netbios	115266	0.02
pplive	5211142	0.04	xunlei	3006009	0.06	napster	91234	0.02
smb	3193585	0.03	qq	1466992	0.03	smb	77664	0.02
napster	1967830	0.02	smb	946538	0.02	soulseek	76431	0.02
yahoo	1916572	0.02	napster	905483	0.02	yahoo	71683	0.01
qq	1721833	0.01	netbios	860923	0.02	pplive	68794	0.01
netbios	1368524	0.01	yahoo	747789	0.02	rtsp	45383	0.01
soulseek	1023863	0.01	soulseek	475708	0.01	qq	29852	0.01
ftp	833635	0.01	ftp	349632	0.01	ftp	27157	0.01
freenet	802869	0.01	x11	147549	0	x11	23707	0
x11	439587	0	armagetron	138125	0	armagetron	14103	0
marca	275851	0	marca	103379	0	rlogin	10190	0
vnc	186206	0	rlogin	62825	0	marca	9885	0
armagetron	181626	0	freenet	59560	0	megaupload	5083	0

imap	155463	0	gnutella	50471	0	elMundo	4320	0
gnutella	100040	0	sip	45292	0	freenet	3816	0
megaupload	96078	0	youTubeClick	35147	0	sip	3508	0
sip	95684	0	megaupload	34878	0	abc	3469	0
youTubeClick	85611	0	imap	30616	0	megauploadA	3104	0
elMundo	69054	0	abc	29492	0	gnutella	2886	0
abc	68173	0	elMundo	25134	0	nntp	1460	0
rlogin	58897	0	megauploadA	16035	0	youTubeClick	1460	0
aim	46404	0	ntp	11714	0	ntp	1215	0
megauploadA	39356	0	nntp	11680	0	aim	588	0
rdp	31489	0	rdp	9299	0	elPais	275	0
ntp	29905	0	elPais	7416	0	irc	179	0
nntp	21907	0	h323	4368	0	shoutcast	146	0
jabber	17495	0	rapidshareAcc	3407	0	imap	51	0
elPais	17225	0	irc	2563	0	lpd	24	0
rapidshareAcc	16988	0	shoutcast	439	0	**TOTAL**	**482704587**	**100**
irc	6002	0	lpd	294	0			
h323	2629	0	aim	270	0			
rapidshare	1396	0	gopher	235	0			
shoutcast	877	0	ventrilo	201	0			
lpd	789	0	jabber	181	0			
gopher	258	0	vnc	168	0			
ventrilo	201	0	imesh	80	0			
imesh	200	0	**TOTAL**	**4822665456**	**100**			
pcanywhere	2	0						
TOTAL	**12058331964**	**100**						

Tabla 4-17: Porcentaje por Aplicaciones Muestreo Distribuido

La comparación de esta métrica se muestra en este estudio para distintas tasas y una única combinación de *offsets*, por sencillez en la presentación. En ella aparecen primero las listas de aplicaciones identificadas y la cantidad de datos correspondientes a esas aplicaciones que han pasado por la red, y una distribución del reparto de esta carga de datos según la división de aplicaciones por grupos.

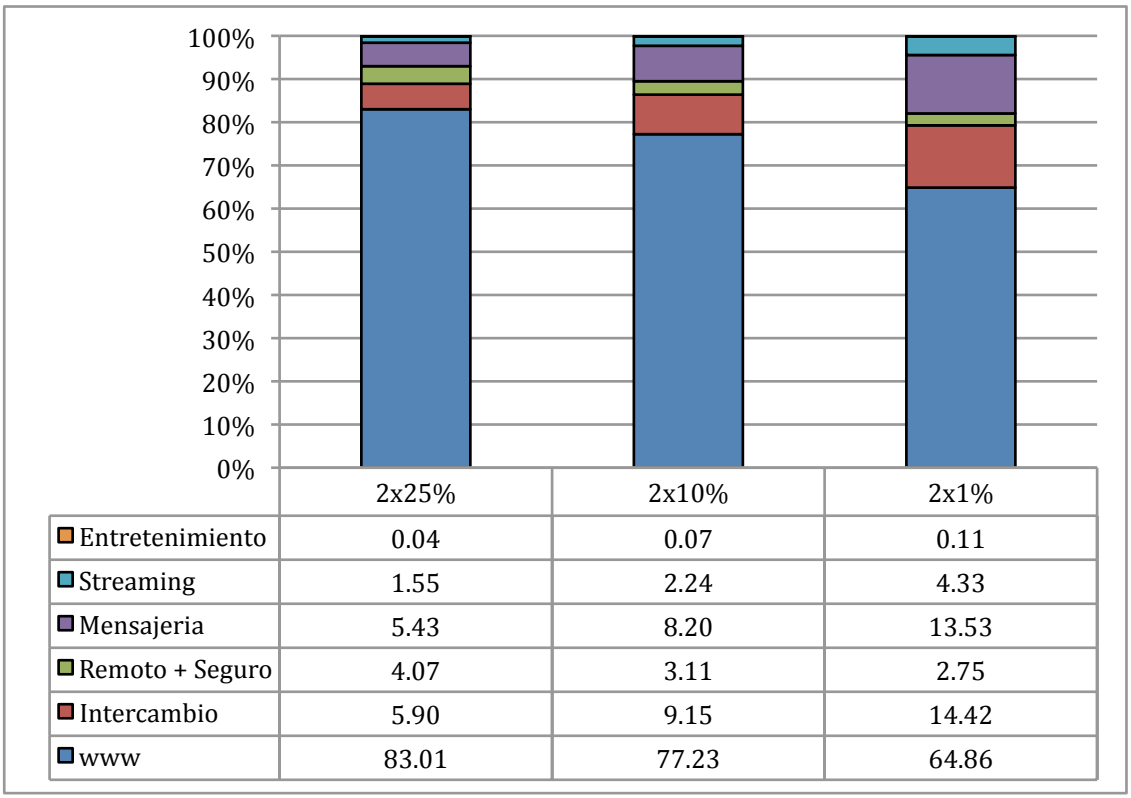


	2x25%	2x10%	2x1%
Otros	51.19	67.34	79.82
Entretenimiento	0.02	0.02	0.02
Streaming	0.76	0.73	0.87
Mensajeria	2.65	2.68	2.73
Remoto + Seguro	1.99	1.02	0.56
Intercambio	2.88	2.99	2.91
www	40.52	25.22	13.09

Ilustración 4-34: Distribución de Tráfico por aplicaciones. Muestreo Distribuido.

En esta primera gráfica se incluye el grupo de aplicaciones no identificadas, englobadas bajo la etiqueta "*Otros*". De nuevo es claro el impacto que tiene la disminución de la tasa de muestreo en la aparición de flujos no identificables. Los valores que se obtienen son siempre mejores que para muestreo simple a la misma tasa, y parecen seguir la tendencia casi lineal que este porcentaje tiene con la tasa de muestreo.

La siguiente gráfica muestra el porcentaje de datos total por grupos, sin tener en cuenta las aplicaciones no identificadas:



	2x25%	2x10%	2x1%
Entretenimiento	0.04	0.07	0.11
Streaming	1.55	2.24	4.33
Mensajeria	5.43	8.20	13.53
Remoto + Seguro	4.07	3.11	2.75
Intercambio	5.90	9.15	14.42
www	83.01	77.23	64.86

Ilustración 4-35: Distribución de Tráfico Identificado. Muestreo Distribuido.

De nuevo, y al igual que para el caso de muestreo simple, los porcentajes relativos de tráfico de la mayor parte de grupos aumenta, en detrimento de "*WWW*" y "*Remoto + Seguro*" . Una posible causa puede ser que las firmas aparecidas para este tipo de tráfico se ven segmentadas y por lo tanto quedan inservibles para su identificación. La evolución de los porcentajes relativos de todos los grupos sigue la línea observada para muestreo simple, por lo que se considera que son resultados perfectamente válidos en comparación con este sistema, y que de nuevo ofrece mejores informes con la misma tasa de muestreo.

4.3.7 Flash Flows

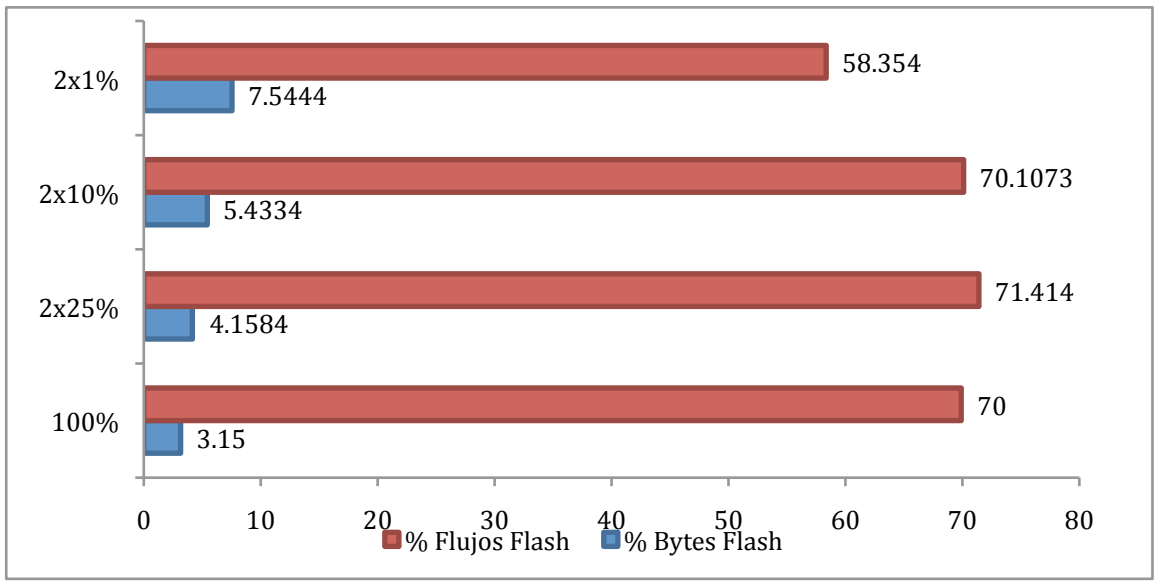


Ilustración 4-36: Flash Flows vs Número de Bytes. Muestreo Distribuido. 15s interarrival

Si se comparan los resultados de muestreo doble al 25% y simple al 50%, se comprueba que la diferencia entre ambos es muy pequeña. Esta comparación con las demás tasas, tanto para muestro doble como simple, muestran también cómo la distribución que sigue esta métrica según las mismas es muy parecida. Esto valida el uso del muestreo doble, que como ya se ha comprobado para esta y otras métricas, proporciona siempre valores cercanos a los obtenibles con muestreo simple a doble valor de tasa, y que aunque en algunos casos minoritarios de valores de *offset* consecutivos este valor no se alcance, siempre es sensiblemente mejor que antes.

De los resultados obtenidos para prácticamente todas las métricas se puede extraer la clara mejora que se obtiene siempre gracias a la agregación de informes distribuidos con muestreo de paquetes. Estos resultados confirman la validez de un sistema de este tipo y aportan valor al trabajo realizado, abriendo la puerta a posteriores investigaciones en este campo.

Otro aspecto a tener el cuenta es el impacto que el propio muestreo, simple o distribuido, tiene sobre las estadísticas obtenidas frente a las generadas con el 100% de la traza original. Según los tests de similitud, y por propia inspección visual de los resultados, la diferenciación estadística con la traza completa es una constante. Este fenómeno promueve la investigación de métodos de monitorización que mejoren las estadísticas obtenidas.

5 Conclusiones y trabajo futuro

5.1 Conclusiones

Una vez mostrados todos los resultados de las métricas extraídas tras las sucesivas ejecuciones del generador de flujos implementado, se considera alcanzado el objetivo inicialmente estipulado para este proyecto de fin de carrera.

El trabajo realizado ha pasado finalmente por los siguientes puntos:

Se ha realizado un estudio exhaustivo del Estado del Arte y de las líneas de investigación relevantes para el desarrollo y estudio de los sistemas de monitorización de red basados en flujos, concretamente en el sistema *netflow*. Este ha abarcado las especificaciones de un sistema de este tipo, y sus particularidades. Se han identificado los problemas actuales que presenta y seguidamente el foco de estudio se ha centrado en las distintas técnicas propuestas últimamente en la literatura científica, con un énfasis en métodos de muestreo, sus parámetros y aplicaciones. Durante este proceso se han identificado también métricas útiles para su posterior análisis, y se han tomado referencias para aplicar a la hora de decidir sobre qué tipo de muestreo implementar y la elección de sus parámetros.

La información adquirida permite tomar una decisión para la implementación de un generador de flujos basado en la definición establecida por *netflow*, con un muestreo por paquetes a distintas tasas y limites máximos. El generador implementado no sólo utiliza métricas simples, sino que también realiza una identificación de las aplicaciones que generan cada flujo detectado mediante distintas técnicas.

Las primeras ejecuciones del generador hacen uso de una traza de datos real de gran tamaño y extraen datos de la misma haciendo un análisis de todos sus paquetes, sin aplicar muestreo. De los informes obtenidos se realiza una extracción de distintas métricas que se utilizan más adelante como referencia. Se hace una **exposición y análisis de los datos** extraídos y se hace una primera valoración sobre las características del tráfico analizado, y sobre el impacto que tiene el tiempo máximo entre llegadas en la definición de flujo. Para ellos se hace uso de tests de similitud basados en la prueba Chi cuadrado de bondad de ajuste. Estos resultados por si mismos ya pueden resultar de interés a la comunidad de

Internet pues posibilita realizar simulaciones o experimentos que utilicen como parámetros los datos reales que en este trabajo se han mostrado.

En la fase siguiente se realiza un gran número de ejecuciones del generador de flujos, esta vez con distintas tasas de muestreo y tiempos máximos entre paquetes. De este largo proceso se extrae información sobre **el impacto** que en efecto tiene este **muestreo** en las estadísticas obtenidas con respecto a la traza completa. Por cada una de las métricas se realiza una valoración exhaustiva de estos efectos, y se trata de encontrar su origen.

Por último, se plantea un escenario de red de monitorización que permita evaluar una técnica de **agregación de informes de flujo** obtenidos en distintos puntos de las misma. Estos informes se realizan mediante muestreo de paquetes. El escenario descrito se centra en un segmento de una red más grande que obtiene una porción concreta del tráfico total mediante identificación de un grupo específico de flujos. Este segmento, que contiene dos equipos generadores de informes, recibe el mismo tráfico, y realiza el muestreo mencionado con un distinto valor de *offset* entre paquetes. Para la simulación de este sistema se implementa una modificación del generador de flujos implementado, que es capaz de realizar ambos muestreos y la agregación de sus informes de manera automática. Para conocer cuál es el impacto que puede tener el *offset* mencionado en la bondad de los informes extraídos, se vuelve a realizar un tren de ejecuciones con distintos valores de tasa de muestreo y offset.

La posterior extracción, por última vez, de las métricas ya consideradas en los casos anteriores, permite de nuevo realizar un análisis pormenorizado de sus resultados en comparación con los de muestreo simple a distintas tasas.

De esta comparación se hace evidente la mejora obtenida mediante muestreo distribuido frente a muestreo simple en la mayoría de los casos. Se comprueba cómo esta mejora planteada puede alcanzar cuantitativamente una bondad equivalente a la obtenible mediante muestreo simple al doble de tasa. De este resultado se extrae que, en el caso de realizar la distribución del muestreo entre más equipos, será bien posible alcanzar los mismos valores de bondad con tasas de muestreo inversamente proporcionales al número de generadores de flujos empleados en la subred de monitorización. También se comprueba cómo la diferencia en el *offset* del muestreo de los paquetes de la traza tiene en

general un impacto muy pequeño en los resultados obtenidos, a excepción de unos pocos casos particulares que corresponden a Offset consecutivos. Los resultados obtenidos para estos casos siguen siendo mejores que para aquellos con la misma tasa a muestreo simple, pero no llegan a ese valor multiplicativo proporcional al número de generadores. De este fenómeno se extrae por lo tanto que, a la hora de una implementación física de un sistema similar, es conveniente coordinar una separación entre Offset de muestreo, que garantice resultados óptimos. Se plantea la consecución de esta coordinación mediante la utilización de banderas en paquetes "clave" que permitan a cada generador calcular el *offset* a aplicar en su muestreo.

Este estudio ha comprobado la posibilidad de la agregación de informes *netflow* obtenidos mediante muestreo de paquetes en distintos puntos de una red de monitorización y la clara mejoría se puede obtener con un sistema de este tipo. Al tratarse de un sistema de implantación sencilla se considera que podría ser una técnica muy útil a tener en cuenta por la comunidad científica y los proveedores de servicios de red.

5.2 Trabajo Futuro

Si bien los resultados obtenidos por este Proyecto se consideran satisfactorios, se identifican posibles líneas de trabajo que amplíen este estudio y puedan proporcionar información sobre otros aspectos del sistema propuesto, no cubiertos por este estudio.

La confirmación de la existencia de "curvas minoritarias" con resultados relativamente peores al resto en el caso de muestreo con *offsets* consecutivos hace necesario la adquisición de un conocimiento más profundo sobre la causa del mismo. Esto, unido a la necesidad de validación de los resultados obtenidos a través de simulaciones con muestreo distribuido que utilicen más de dos nodos de generación de informes promocionan la ampliación de este estudio con estas ejecuciones, que permitan confirmar ambos puntos. Su realización se considera sencilla en cuanto a implementación si se basa en el trabajo aquí realizado y fue de hecho un punto considerado durante la realización de este proyecto, pero fue finalmente descartada por considerarse excesiva en cuanto a la considerable extensión de tiempo que implicaría su finalización, y por su relativa redundancia con respecto a los resultado ya expuestos.

Otra línea que parte directamente de los resultados obtenidos es la implementación del sistema propuesto en un equipo real de red, que permita comprobar cuáles son las restricciones efectivas de tiempo que el muestreo de paquetes distribuido y la agregación de sus informes presenta. De un estudio de esta naturaleza se pueden extraer con seguridad conclusiones en cuanto a especificaciones necesarias para equipos basados en este sistema y que se pretendan desplegar en una red comercial, amén de proporcionar una validación definitiva sobre los resultados aquí expuestos.

Por último, la simulación realizada se basa en el sistema planteado por [16], que hace una selección previa de los flujos a muestrear por cada subconjunto de la red total de monitorización. Según este estudio, el sistema es relativamente robusto a cambios en tendencias de tráfico, y su cobertura en cuanto a la captura relativa de todos los flujos asignados es buena. El mismo documento, sin embargo, plantea una extensión a su trabajo que confirme esta característica frente a cambios en las tablas de rutado. La implementación en equipos reales antes planteada podría ser de nuevo la herramienta que valide estos resultados y que siente finalmente las bases de un sistema comercial desplegable basado en esta técnica de muestreo distribuido.

6 Referencias

[1] [N. Browne, Traffic Flow Measurement: Experiences with NetTraMet, IETF RFC2123, Mar 1997]

[2] [Luca Deri, "Ntop", http://www.ntop.org/]

[3] [Se-Hee Han, Myung-Sup Kim, Hong-Taek Ju, J.W. Hong, The Architecture of NG-MON: A Passive Network Monitoring System, In Proceedings of DSOM 2002, Montreal, Canada, Oct 2002 p. 16-27]

[4] [Daniel W. McRobb, "cflowd design", CAIDA, Sep. 1998]

[5] [Internet Protocol Flow Information eXport, http://datatracker.ietf.org/wg/ipfix/charter/]

[6] [Baek-Young Choi, Supratik Bhattacharyya, "Analysis of Traffic Monitoring via Sampled NetFlow ", In ACM SIGMETRICS Performance Evaluation Review Vol 33, Issue 3, 2005,pp 18-23]

[7] ["Cisco Sampled NetFlow", http://www.cisco.com/]

[8] [Gianlucca Iannaccone, Intel Research Cambridge, "CoMo: An Open Infrastructure for Network Monitoring",http://como.intel-research.net/pubs/como.agenda.pdf]

[9] [J. Oberheide, M. Go , and M. Karir. Flamingo: Visualizing internet traffic. In Proceedings of . NOMS 2006, Vancouver, Canada, pp 150-161.]

[10] [Myung-Sup Kim, Young J. Won, James W. Hong, "Characteristic Analysis of Internet Traffic from the Perspective of Flows", In Journal of Computer Communications, Vol 29, Issue 10, June 19 2006, pp 1639-1652]

[11] [cSamp: A System for Network-Wide Flow Monitoring'', Vyas Sekar, Michael K. Reiter, Walter Willinger, Hui Zhang, Ramana Rao Kompella, David G. Andersen. In Proceedings of 5th USENIX NSDI, San Francisco, CA, Apr., 2008]

[12] [Baek-Young Choi, Supratik Bhattacharyya, "Observations on Cisco Sampled Netflow", ACM SIGMETRICS Performance Evaluation Review - Special issue on the First ACM SIGMETRICS Workshop on Large Scale Network Inference (LSNI 2005) Volume 33 Issue 3, December 2005]

[13] [Kimberly C. Claffy, George C. Polyzos. Hans-Werner Braun, "Application of Sampling Methodologies of Network Traffic Characterization" in Proceeding SIGCOMM Conference proceedings on Communications architectures, protocols and applications, San Francisco, CA, 2003.]

[14] [Edurne Izkue, Eduardo Magaña, "Sampling Time-Dependent Parameters in High-Speed Network Monitoring", In Proceedings of the ACM international workshop on Performance monitoring, measurement, and evaluation of heterogeneous wireless and wired networks, Málaga, Spain, 2006.]

[15] [Nicolas Hohn, Darryl Veitch, "Inverting Sampled Traffic" In IEEE/ACM Transactions on Networking (TON) Volume 14 Issue 1, February 2006]

[16] [W. Vyas Sekar, Michael K. Reiter, Walter Willinger, Hui Zhang, "Coordinated Sampling: An Efficient, Network-Wide Approach for Flow Monitoring", Technical Report,CMU-CS-07-139, Computer Science Dept, Carnegie Mellon University, 2007]

[17] [W. J. Lui, j. Gong, "Double Sampling for Flow Measurement on High Speed Links", In Computer Networks Volume 52 Issue 11, August, 2008]

[18] [Internet 2 Weekly Reports, http://netflow.internet2.edu/weekly, 2010]

[19] [C. Estan, G. Varghese, "New Directions in Traffic Measuremtent and Accounting: Focusing on the Elephants, Ignoring the Mice", ACM Transactions on Computer Systems (TOCS) Volume 21 Issue 3, August 2003]

[20] [IETF psamp WG, http://datatracker.ietf.org/wg/psamp/charter/]

[21] [Daniela Brauckhof, Brenhard Tellenbach, Arno Wagner, Martin May, Anukool Lakhina, "Impact of Packet Sampling on Anomaly Detection Metrics", In Proceedings of the ACM SIGCOMM conference on Internet measurement, Rio de Janeiro, Brazil, 2006.]

[22] [Wireshark, http://www.wireshark.org/]

[23] [Libpcap, http://www.tcpdump.org/]

[24] [Alejandro López Monge, "Aprendiendo a programar con Libpcap", http://www.e-ghost.deusto.es/docs/2005/conferencias/pcap.pdf]

[25] [MacFuse, http://code.google.com/p/macfuse/]

[26] [MacFusion, http://macfusionapp.org/]

[27] ["L-7 filter supported protocols", http://l7-filter.sourceforge.net/protocols]

[28] ["Toward the accurate identification of network applications", Moore, A.W., Papagiannaki, K., In Proceedings of Passive & Active MeasurementWorkshop 2005, Boston, MA, 2005.]

[29] [Gianlucca Iannaccone, Christophe Diot, Ian Graham, Nick McKeown, "Monitoring Very High Speed Links", In Proceedings of the ACM SIGCOMM Workshop on Internet Measurement, San Francisco, CA, 2001.]

[30] ["Web User-Session Inference by Means of Clustering Techniques",Andrea Bianco, Gianluca Mardente, Marco Mellia, Maurizio M. Munafò, Luca Muscariello. In IEEE/ACM Transactions on Networking, Vol.17, Issue 2, pp.405-416,]

[31] [Nick Duffield, Carsten Lund, Mikkel Thorupp, "Estimating Flow Distributions From Sampled Flow Statistics", In IEEE/ACM Transactions on Networking (TON) Volume 13 Issue 5, October 2005]

[32] [Nick Duffield, Carsten Lund, Mikkel Thorup, “Properties and Prediction of Flow Statistics from Sampled Packet Streams”, IEEE/ACM Transactions on Networking (TON) Volume 13 Issue 5, October 2005]

[33] [Cristian Estan, Ken Keys, David Moore, George Varghese, ”Building a Better Netflow”, In Proceedings of the Conference on Applications, technologies, architectures, and protocols for computer communications, Portland, OR, 2004.]

[34] [“Introduction to Cisco IOS NetFlow - A Technical Overview”, http://www.cisco.com]

[35] [“CoralReef – CAIDA”.http://www.caida.org/tools/measurement/coralreef/]

[36] [Gonzalo Polo Vera, “Caracterización de Tráfico a Partir de Trazas de Trafico Reales: Aplicaciόm a RedIRIS”, PFC UAM]

[37] [C. Estan, G. Varghese, “New Directions in Traffic Measurement and Accounting”, In ACM Transactions on Computer Systems TOCS Volume 21 Issue 3, August 2003]

[38] [Mark E. Crovella, Murad S. Taqqu and Azer Bestavros, “Heavy-Tailed Probability Distributions in the World Wide Web”, A practical guide to heavy tailsBirkhauser Boston Inc. Cambridge, MA, 1998]

PRESUPUESTO

1) Ejecución Material

- Compra de ordenador personal ..... 1.500 €
- Alquiler de servidor durante 1 año ..... 1.000 €
- Compra de Disco duro portátil. Capacidad 2TB ..... 100 €
- Material de oficina..... 50 €
- Total de ejecución material ..... 2.650 €

2) Gastos generales

- 16 % sobre Ejecución Material ..... 424 €

3) Beneficio Industrial

- 6 % sobre Ejecución Material ..... 159 €

4) Honorarios Proyecto

- 1280 horas a 15 € / hora ..... 19.200 €

5) Material fungible

- Gastos de impresión ..... 50 €
- Encuadernación ..... 50 €

6) Subtotal del presupuesto

- Subtotal Presupuesto ..... 22.533 €

7) I.V.A. aplicable

- 16% Subtotal Presupuesto..... 3.605,32 €

8) Total presupuesto

- Total Presupuesto ..... 26.138,32 €

Madrid, Junio de 2011

El Ingeniero Jefe de Proyecto:
José Luis García Dorado, PhD

Fdo.: Fernando Gutiérrez de Rubalcava Blanca
Ingeniero Superior de Telecomunicación

PLIEGO DE CONDICIONES

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de unsitema de monitorización de red basado en *netflows* mediante muestreo distribuido de paquetes. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es

obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad "Presupuesto de Ejecución de Contrata" y anteriormente llamado "Presupuesto de Ejecución Material" que hoy designa otro concepto.

Condiciones particulares

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.